uc3m | Universidad Carlos III de Madrid

Degree in Telecommunication Technologies Engineering
Academic Year 2018-2019

Bachelor Thesis

"Modified Band Depth Based Initialization of K-Means for Functional Data Clustering"

Javier Albert Smet

Tutor

Aurora Torrente Orihuela

Leganés, 2019

SUMMARY

K-Means is a well-known clustering algorithm that produces optimal results with a correct initialization. Modified Band Depth rises as a reliable alternative to one of the most widespread initialization methods: K-Means Plus Plus. Through the B-Spline approximation of the observations of interest, multivariate outcomes can be interpreted as functional data to produce a better grouping than via direct clustering. Several models and real data are used to determine if the method proposed produces favorable results in a consistent manner. We find that our method works well and outperforms the alternatives in most of the situations, specially achieving better clustering accuracy. This project is the continuation of the research work done in the Statistics Department of UC3M on Modified Band Depth.

Keywords

Clustering; K-Means; Modified Band Depth; Functional Data; B-Splines; Bootstrapping

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Prof. Aurora Torrente, for trusting and supporting me throughout my undergraduate years, and for her motivation and guidance in this project. I cannot conceive a better tutor than the one I have had, especially centered in my learning and personal growth.

Besides my advisor, I would like to thank my father, for his time and perpetual patience, my mother, for her reliable advice, and my brother, for his inexorable energy.

So far, my entire academic career has been made possible thanks to all the teachers and professors that have believed in my talent and competence. Their guiding knowledge will always shine in my path.

To the International Relations and Cooperation Department of UC3M, for making the world smaller and taking me to Georgia Tech and UNSW, giving me the chance to broaden my mind and expand my heart in the best universities in the world.

TABLE OF CONTENTS

LIST OF ACRONYMS

ARI	Adjusted Rand Index
AWGN	Additive White Gaussian Noise
CoPADIT	Correctness, Purity, ARI, Distortion, Iterations and Time
DF	Degrees of Freedom
EM	Expectation Maximization
Eq.	Equation
FDA	Functional Data Analysis
Fig.	Figure
FMBD	Functional Modified Band Depth
FKMPP	Functional K-Means Plus Plus
KM	K-Means
KMPP	K-Mean Plus Plus
MBD	Modified Band Depth
MVMBD	Multivariate Modified Band Depth
OSF	Oversampling Factor
PAM	Partitioning Around the Medoids
PERT	Program Evaluation and Review Technique

FIGURE INDEX

TABLE INDEX

EQUATION INDEX

1. INTRODUCTION

1.1 Problem Definition

The field of data analysis has become increasingly popular due to the ease of data collection allowed by current technology. Nowadays we can extract trends and statistics which can be analyzed and interpreted to reach meaningful conclusions in every research field.

Collected data can represent different phenomena and can be mined by using various tools like sensors or questionnaires. The latter ones are usually identified with discrete, multivariate data, while the former ones tend to provide continuous multivariate data or functional data.

To illustrate this, let us take the example of an athlete. Questionnaires about his/her health, like sleeping or eating habits, would provide us with some categorical data; his height, weight or his best performance in a 100m race are continuous variables that record continuous features that cannot be fit together into a function. However, if we measure the athlete's blood pressure every half a minute of the race, we can figure out that the data collected are just observations of a function 30 seconds apart. This means that any value between two given observations exists (i.e.: the athlete has a specific blood pressure in between the time instants that we measure), but we just have not observed it.

Functional data analysis has multiple applications, such as monitoring patients at hospitals [1], statistically interpreting electromagnetic signals [2], or classifying children growth patterns [3].

Among the vast collection of methods to analyze data, clustering will be the focus of this work. Clustering is a form of unsupervised classification [4] that serves to infer groups of individuals or objects with similar characteristics and has a broad range of applications.

1.2 Motivation

The uses of clustering vary widely depending on the field [5]. Identifying diseases in medicine [6], [7], grouping clients by purchase habits in marketing [8], or noise removal in signal processing [9] are just some examples of it.

Improving the performance of a clustering algorithm translates into a more effective disease treatment, increased profit of an organization due to better customer targeting or more accurate signal processing results.

Generally speaking, this research project, although quite technical and specific, is relevant in the data analysis ecosystem, a field with many applications that can be life-changing through its uses in medicine, biology, market analysis or education amongst others.

1.3 State of the Art

Clustering algorithms have formally been in place since the mid-twentieth century. There are several approaches that have been developed up-to-date and will be described in detail in the *Methods* section. Amongst these, the most extended non-hierarchical clustering algorithm used for functional data analysis is K-Means [10].

It is known that K-Means converges quickly to a local optimum [11], but fails to provide consistent results when initialized randomly. The key for getting an optimal result is to carefully provide a set of initial centers as initialization parameters.

The finite dimensional version of the Modified Band Depth (MBD) [12], a concept that generalizes univariate medians to higher dimensions, is presented in Torrente and Romo [13] as a key element to obtain a solution to this initialization problem, providing better results than other used initialization algorithms for multivariate data. However, up to now, this algorithm has not been tested for functional data.

1.4 Goals

A Final Year Project can be understood as a double opportunity, both personal and scientific. It is a chance for the author to gain a thorough insight into a degree-related topic, and a chance to produce a meaningful research piece of work. From the research perspective, saving the more personal insight for the discussion section, the main purpose of this project is to assess whether an MBD-based initialization of K-Means is convenient for clustering of functional data. To achieve this, the fundamental lines of work can be broken down into:

1. Addressing the transformation of input multivariate data into functions for clustering analysis.
2. Assessing the clustering output results using a set of performance evaluation measures and comparing the proposed method to other initialization methods.
3. Determining whether MBD-based initialization is an advantageous solution for K-Means clustering in the case of functional data.

Additionally, although the project involves many tests and calculations, we have made an effort to make the document clear, visually-attractive and as easy-to-read as possible, without sacrificing scientific correctness and conciseness.

2. PLANNING

2.1 Initial Planning

The project has an estimated duration of 34 weeks - or equivalently, 8 months - from start to finish. The tasks to be completed are summarized in Table 2.1.

TABLE 2.1.
TASK BREAKDOWN AND DURATION.
The following abbreviations are used: I = Introduction, D = Development, C = Closing

Reference	Task	Duration (weeks)
I1	Literature review and state of the art familiarization	4
I2	Approach to the R language, adaptation to R-Studio and selection of packages to be used	4
D1.1	B-Spline approximation implementation	3
D1.2	Application of K-Means after a B-Spline approximation	1
D2.1	Definition of the models to be tested	3
D2.2	Parameter optimization for the models	3
D2.3	Model testing	3
D2.4	Study of missing data	2
D3.1	B-spline coefficient clustering	3
D4.1	Comparison with other initialization methods	2
C1	Future research line considerations	2
C2	Writing of the final document	4
TOTAL		34

Note that the table has necessarily been updated with the specific details of the tasks that were not known exactly from the start. However, the main outline associated to the different letters of the reference column has remained unchanged, which is represented Fig. 2.1.

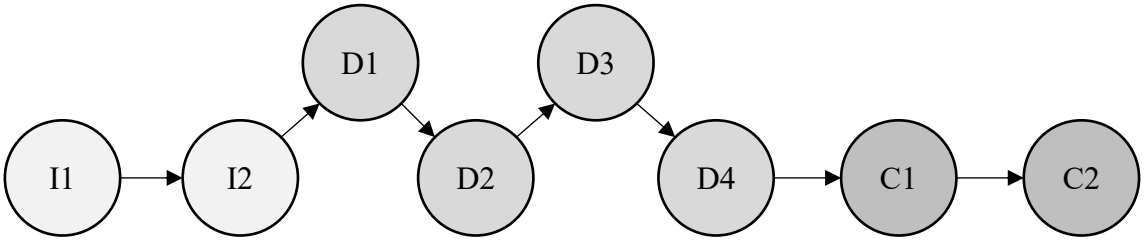


Fig. 2.1. PERT diagram describing the tasks carried out in this project.

The diagram presents only one path from start to finish as all the tasks to be done are sequential. In other words, there is no way in which we can perform two tasks at the same time. Hence, the critical path is the only one to follow, and all activities are therefore critical activities (i.e. they must be completed on time).

2.2 Regulatory Framework

As it happens frequently in the Computer Science branch of Telecommunications, there is a lack of regulation in the matter. Considering that there is nothing physical to be built and algorithms cannot be patented as such [14], the regulatory framework is reduced to one main legal issue: privacy, the biggest legal concern when talking about Data Science.

Particularly, the algorithm proposed has to be tested on real datasets. These have been obtained from sites that make them publicly available, and that are in compliance with the data privacy regulations established by the Regulation (EU) 2016/679 of the European Parliament [15].

2.3 Choice of Programming Language (R) and Software (R-Studio)

Having set up the backbone of what the project involves, it is important to explain the reasoning behind the choice of software in which it is developed. Why R? Why R-Studio?

R is the most popular programming language amongst statisticians [16]. It is intuitive and flexible. Furthermore, it is open-source: there is a huge community of contributors, who provide code and documentation that can be imported as packages to ease the programming tasks.

R-Studio is a free integrated development environment similar to Visual Studio for C and C++, or Eclipse for Java. It provides the programmer with a graphical user interface that aids the visualization of the workspace: all the variables and functions that are in memory can be intuitively seen, as well as the console window. It is easy to write scripts and run them, and to debug the code line by line. R-Studio is used for research in top-class universities all around the world [17], [18].

3. METHODS

In this section the research method is described. The whole picture can be better understood if it is thought of as being made up of three components:

1. A literature review (subsection 3.1) to cover the main concepts involved in clustering and functional data, and to understand why we are researching specifically in K-Means initialization.
2. A case study / research project (subsections 3.2 to 3.6) to assess if MBD-based initialization of K-Means is an accurate and reliable solution for centroid-based clustering.
3. The development of a computer program (subsection 3.7) that takes some input data and outputs the cluster each datapoint belongs to by using MBD-based initialization of K-Means.

Although every approach will be explained in more detail in the subsections mentioned, Fig. 3.1 presents the overall outline of what will be covered, understanding the initialization algorithm proposed as a system.

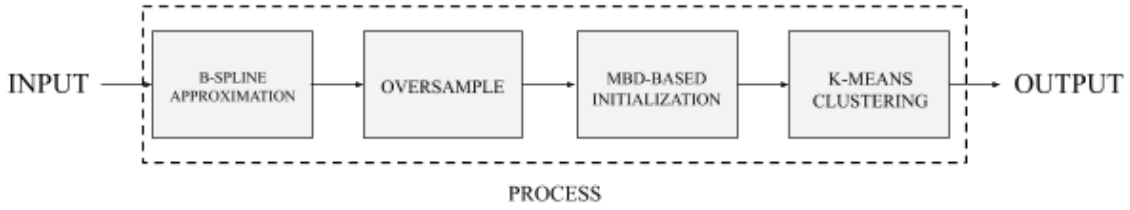


Fig. 3.1. The algorithm as a system. Input data is provided and transformed during the different steps of the method to render an appropriate output.

This is nothing but the classic picture of a system with an input, some processing and an output. It is handy to be familiar with this figure so we do not lose the grip on the goals of the project. It is important to note that the processing unit associated to "MBD-BASED INITIALIZATION" makes up the innovative block of the system. Table 3.1 presents a summary of the concepts used.

TABLE 3.1.
PARTS OF THE SYSTEM.

Part of the system	Role in the project
Input	Datapoints that will be treated as functional data.
Process	Core of the research work; B-Spline function fitting and MBD-based initialization of K-Means.
Output	A label for each input datapoint (function) indicating which cluster it belongs to.

For simplicity, the whole processing block of the system is referred to as Functional Modified Band Depth, or FMBD, short for Functional data clustering based on Modified Band Depth initialization of K-Means.

3.1 Literature Review

The literature consulted in task I1 of the initial planning is summarized in this section. The theoretical foundations to understand the project in depth are laid here. The practical application of this theory is then explained in section 3.2 (*Technique Review*). A personal recommendation for the most experienced readers is to jump into the subsections that are unfamiliar to them.

3.1.1 Multivariate and Functional Data

What is meant by functional data is best understood through an example.

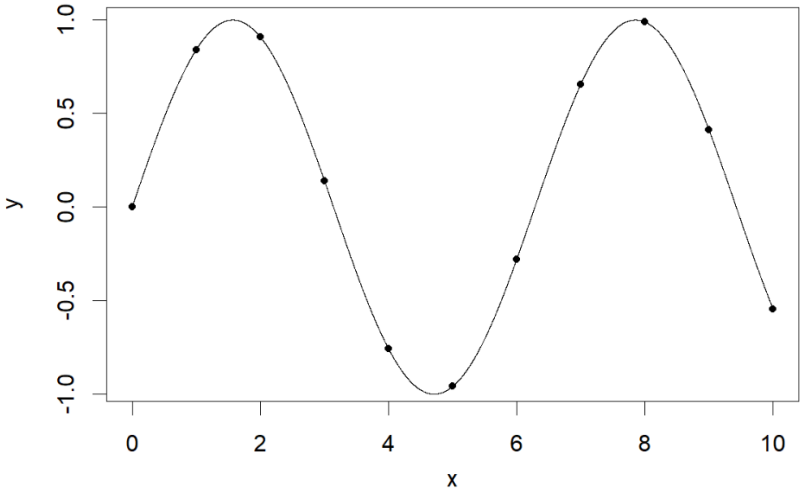


Fig. 3.2. Function sampling. A continuous function (black line) is observed at discrete x values (dots).

The function $y = sin(x)$ is plotted in Fig. 3.2. The black dots, drawn on top of the graph of the function, represent points on the plane, with x and y coordinates, which are obtained just by assessing the function at specific values of x. For example, for $x=0$, we have $y = sin(0) = 0$, which is the first point in the graph. These dots are called *observations or samples* of the function, and the assembly of them is said to be a functional *datum* or *datapoint*. Now let us consider the setup in Fig. 3.3:

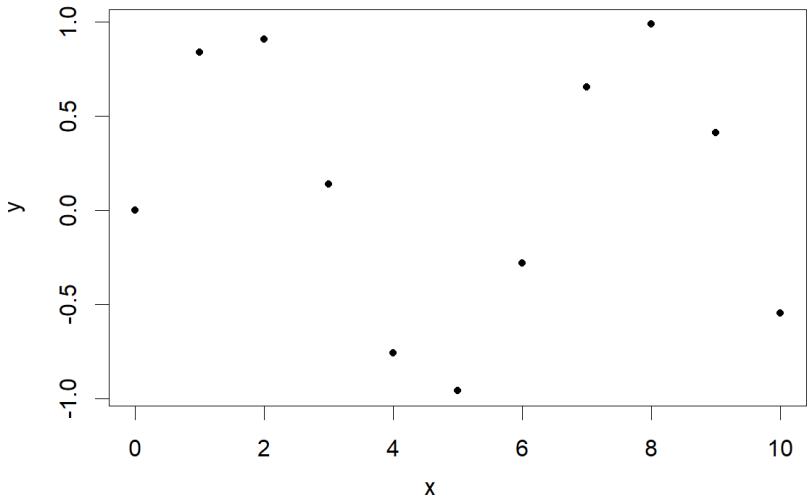


Fig. 3.3. A functional datum as a collection of points on the plane.

The points on the plane are the same but now the relationship with the underlying function is not as obvious as before. It is even less clear if we reduce the number of observations we have. Therefore, the more points we observe, the easier it is to reconstruct the original function [19]. Having a large amount of observations in comparison to the duration of the function (i.e. the length of the interval in which the function is observed, 10 in this case) means that we have a high *sampling rate*.

Note that all of the concepts described here, like samples, observations and sampling rates are related to *functions*.

Another important aspect that characterizes a collection of data points as a functional datum is the fact that there is an ordering in the samples. That is, the observation at $x = 0$, comes before the sample at $x = 1$, which comes before the one at $x = 2$, and so on. Furthermore, observations are considered to be in a continuum, so that we accept that any value between $x = 0$ and $x = 1$ exists, but we just have not observed it. In particular, the x-axis is usually understood as the *time* axis.

Moreover, functional data analysis is very flexible, in the sense that functions can be observed at time points that are not equally spaced or that can vary across functions. Nevertheless, it is not unusual to find many functions observed for the same set of values of x, that is, observed at the same time points.

However, there is another type of data that, in general, does not originally come from a function. Instead, a datum is simply a vector of two or more *variables* or *features*. This is what we refer to as *multivariate* data.

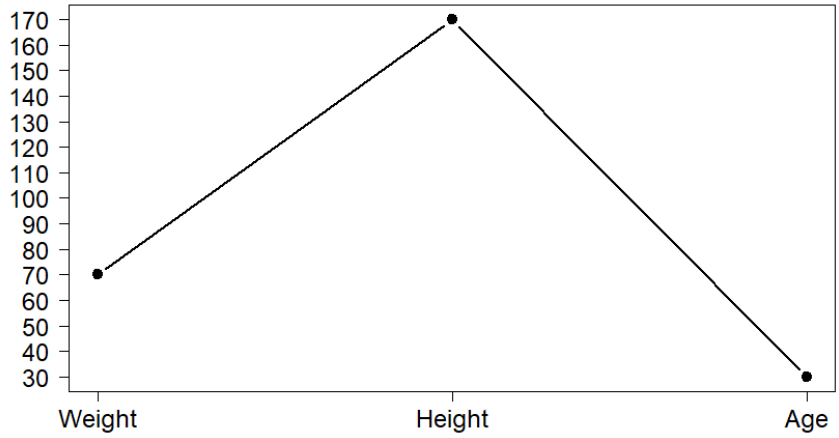


Fig. 3.4. Multivariate data representing different features of an individual.

In the example proposed in Fig 3.4, the information for a fictional person has been collected. His weight, height and age are plotted in the same graph. The weight is not measured in the same units as the height, nor the age, yet we can still plot this information in the same graph and join them with a line to get the corresponding parallel coordinates of the multivariate vector, like in Fig. 3.4.

In contrast to functional data, multivariate data are not sequential. This means that we can change the order of height and weight in the graph and still have the same information (it is permutation invariant). Similarly, we do not have any values between the points in the graph as we had with the functional data case. For example, there is no value between "weight" and "height" (but there were – unobserved – values between $x = 0$ and $x = 1$ in Fig. 3.3).

Fig. 3.5 shows an example of the re-ordered multivariate data, with the same information.

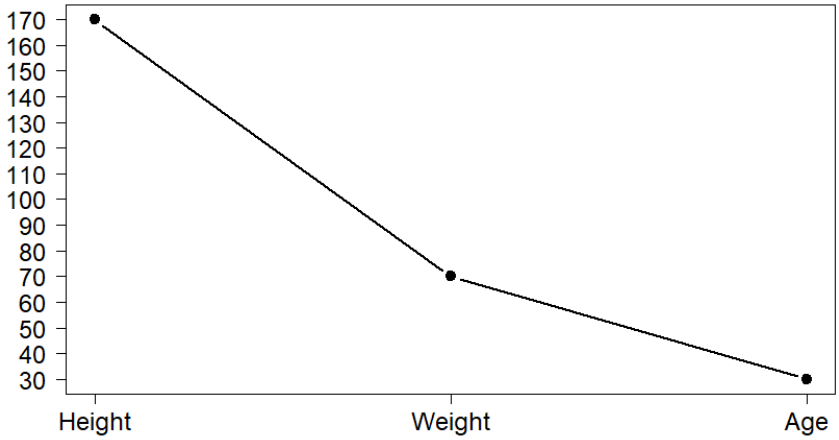


Fig. 3.5. Reordered multivariate data. The information provided by the datapoint is permutation invariant.

As one can envisage, the analysis of data arising from a function as multivariate data reduces the dimensionality of the problem but completely ignores crucial information such as smoothness of the curves or correlation of observations.

An important question arises when telling apart multivariate from functional data. In this project it is *assumed* that the captured data comes from a function and it is treated as such[1].

The simulated models that we have chosen in this work, and that will be explained in section 3.3 (*Models for Testing*), are all made up of functional data. Also, in real life situations, we sometimes come across observed data that are derived from a function, representing a signal, a growth curve or other physical phenomena that can be thought of as functions.

3.1.2 Function Approximation Methods

Once we have our observations of a function, which we will consider our input data, we have to choose a function that is suitable to represent them. As we do not know the exact original function that the data comes from, we have to make a guess.

Of course, we have to define a more formal way of obtaining a function that represents the data than just "guessing". The first idea that comes to mind to solve this problem is the use of models.

Model Selection

Imagine we have a set of functions from which we know the data has been observed. Typically, in the problem of model selection there is a collection of potential models. First, the observations have to be fit to each candidate model, often by estimating some parameters, and next the most appropriate one is selected. For an overview on model selection techniques, refer to Ding, Tarokh and Yang (2018) [20].

[1] Actually, we can understand functional data as N-dimensional multivariate data, as we have N observations from a function collected into a vector, with one dimension for each sample.

The simplest situation is as shown in Fig. 3.6. The graph on the left represents a functional datum, and the one on the right displays a set of functions that it could come from. By visual inspection we can tell that the datum comes from the red model and not the green one. This approach is particularly useful when we *know* the set of functions that the data can come from. Nevertheless, this model set is not usually known, so a more flexible approach that allows us to find previously unknown functions for any kind of data has to be considered.

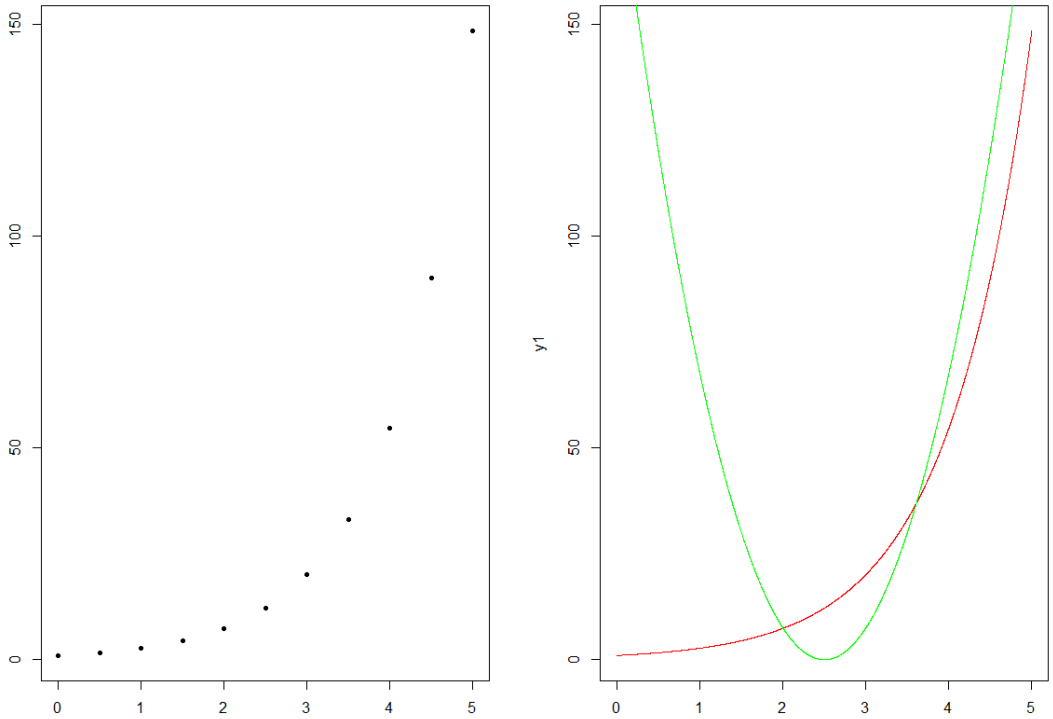


Fig. 3.6. Model fitting. Data points on the left panel are *known* to come from one of the functions on the right.

We now focus on two well-known techniques for representing and approximating functions: Fourier series and B-splines, which are of interest here because they are used for *curve fitting*.

I. Fourier Series

Without going into much detail, Fourier analysis provides us with some *coefficients* that allow representing any periodic function as a linear combination of sine and cosine terms. Fig. 3.7 illustrates the following example, in which the black function corresponds to the following equation:

$$f(x) = 0.5 + 0.6366 \cdot \cos(\pi/2) - 0.2122 \cdot \cos(3\pi/2) + 0.1273 \cdot \cos(5\pi/2). \qquad \text{(Eq. 1.1)}$$

We can see by visual inspection that $f(x)$ approximates the square wave, our target, represented in red, but such an approximation can be drastically improved. The more *terms* or coefficients we add the better the approximation will be. There are well-known closed formulas for each coefficient, namely given by:

$$f(x) = \sum_{n=1}^{\infty} C_n \, e^{j\frac{2\pi \cdot n}{T} \cdot t}, \qquad C_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) \cdot e^{-2\pi j \frac{n}{T} t} dt$$

For a deeper insight into Fourier series Oppenheim, Willsky and Nawab (2014) [21] is recommended.

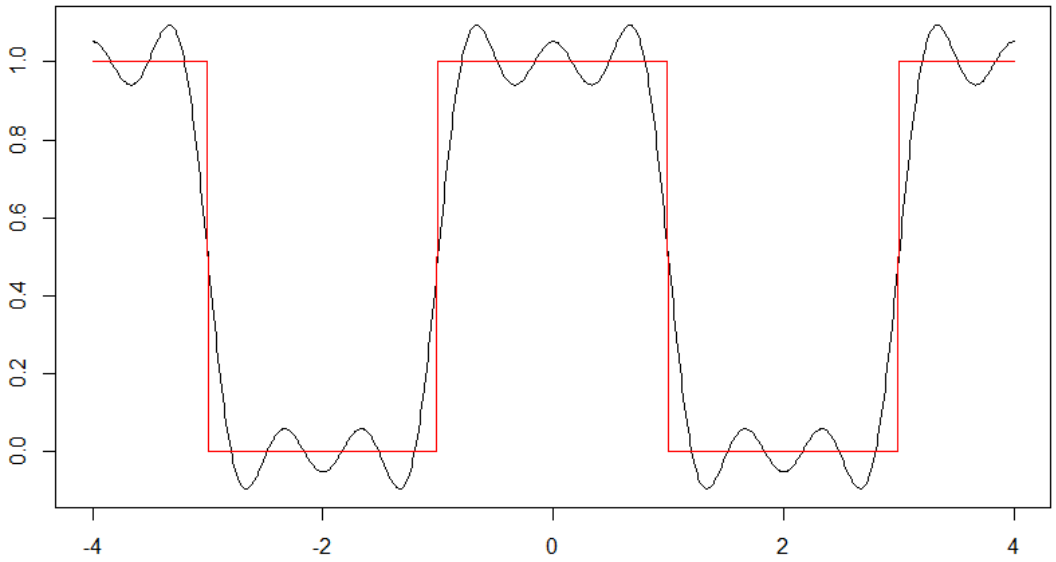


Fig. 3.7. Fourier series approximation (black line) of a square wave (red line).

This example serves to illustrate how the red square wave function can be represented with a reasonably low number of coefficients. From an algebraic perspective, we can represent an *approximation* of the function with two vectors:

- The coefficients vector, [0.5, 0.*6366*, -*0.2122*, *0.1273*],
- and the associated basis vector, [*1*, *cos(π/2), cos(3π/2), cos(5π/2)*].

Doing the dot product between both of them yields the function *f(x)* in Eq. 1.1.

This vector and matrix notation is useful for coding purposes, as R works with such objects.

The same philosophy behind the coefficient-basis notation for representing periodic functions described above applies to B-splines for non-periodic functions. As stated below, B-splines provide other advantages when considering derivatives or efficiency. All these characteristics will be summed up in Table 3.4 at the end of this section.

II. B-splines

Splines are polynomial curves defined *piecewise*. In each piece, or interval of the x axis, they have a specific degree. Algebraically, they can be thought of as elements of a vector space in an interval defined by:

- The degree of the polynomials in a piece, and
- The number of knots, or endpoints of the intervals (pieces).

Moreover, the dimension of the vector space is given by the *degree + 1 + the number of knots*. The *degree + 1* is known as the *order* of the polynomial (the number of terms when all of them are present).

Fig. 3.8 shows an example of a quadratic polynomial defined in the interval [0, 1) and a linear polynomial in the interval [1, 2]. This is a spline of degree 2, with 1 knot plus the two extremes. The knot and the boundary knots are represented as red points. As explained before:

- Order = degree + 1 = 2 + 1 = 3.
- Dimension of the vector space: order + number of internal knots = 3 + 1 = 4.

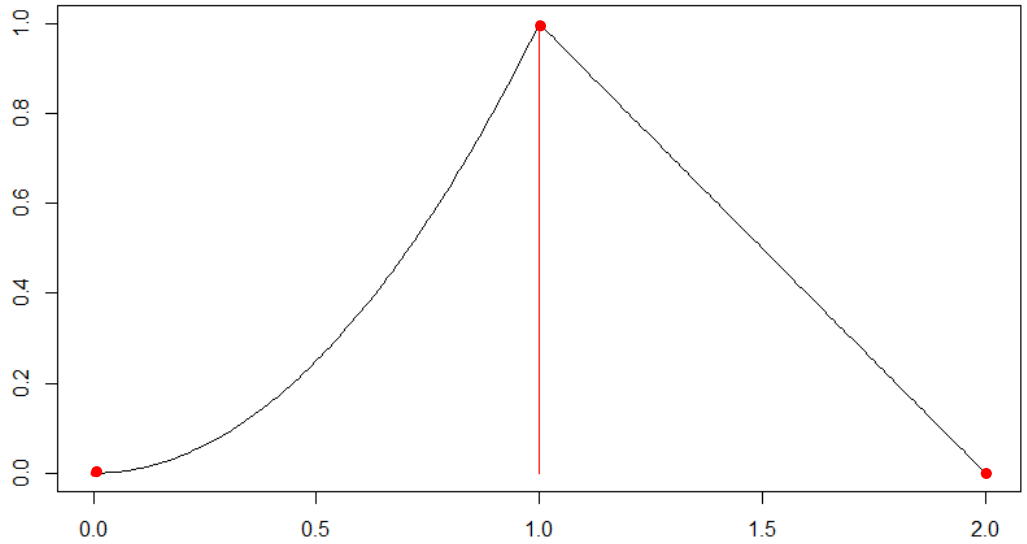


Fig. 3.8. Example of a quadratic spline with one internal knot.

The dimension of a vector space states the number of linearly independent vectors to be in a basis for that space; in the case of splines these basis functions are splines of a specific, common order, and are referred to as B-splines. Because they form a basis, they span all the splines (vectors) of the vector space defined by a specified set of knots and a given order.

One possible, standard basis can be obtained by working recursively from order 1 B-splines, up to the desired order, by using the expression [22]:

$$(m-1)\cdot B_{j,m}(t) = (t-j)\cdot B_{j,m-1}(t) + (m+j-t)\cdot B_{j+1,m-1}(t),$$

where $B_{j,m}(t)$ is the spline basis function of order m and j is the value of t where it first rises from zero. $B_0(t)$ is the spline known as the mother basis spline, a square pulse in the interval [0, 1] with height of 1.

We display in Fig. 3.9 the process of constructing B-splines of order two (B_2), starting from those of order zero (B_0).

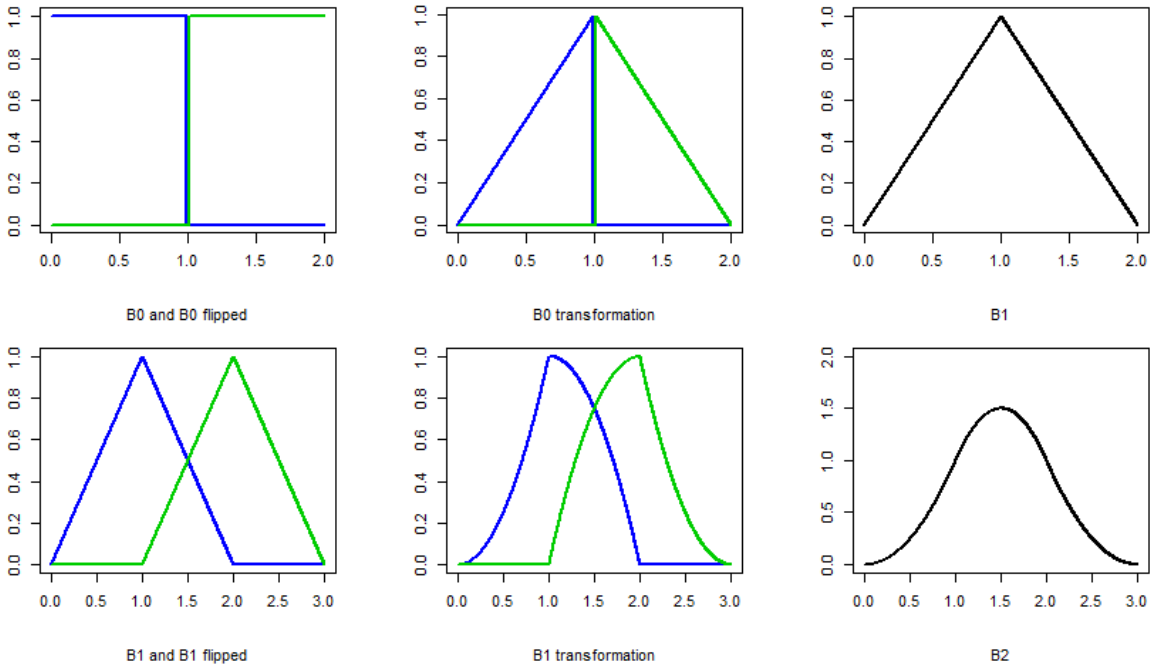


Fig. 3.9. Recursive construction of B-Splines of order 2.

The term B-Spline is shorthand for *Basis*-Spline. This is because we can understand a B-spline as an element of a basis for the vector space of a certain collection of splines.

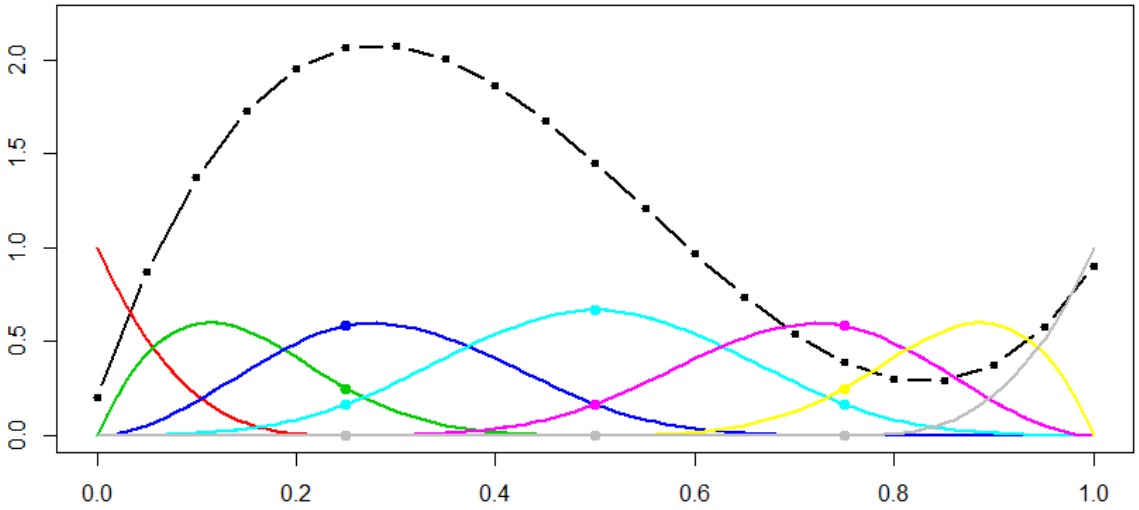


Fig. 3.10. B-splines to span splines with three knots at 0.25, 0.5 and 0.75. The basis consists of 7 cubic splines (colored lines).

Fig. 3.10 serves as an explanation for what a spline is (dash-dotted black curve), as an element of a 7-dimensional vector. The 7 B-Splines in the basis are illustrated as solid, colored lines. The three breakpoints are displayed as dots. The spline is then represented with respect to the given basis by the coefficients (or coordinates) reported in Table 3.2.

TABLE 3.2.
B-SPLINE WEIGHTS.

1	2	3	4	5	6	7
0.2	1.4583	2.475	1.525	0.125	0.2417	0.9

Loosely speaking, *breakpoint* is another name for *knot*, and it refers to the endpoints of the subintervals in which the splines are defined.

Obviously, the dimension of the vector space, as calculated before, gives the number of coefficients we need to represent a whole spline using B-splines.

One major advantage of splines is that they can be used to approximate functions of any kind, not just polynomic functions. B-Splines offer a convenient representation of a function because of the following main characteristics:

- ✓ They have a simple vector representation of a function and its derivatives of any order.
- ✓ A fixed basis can be used for some given knots and order.
- ✓ Basis functions are continuous for any order and differentiable for orders higher than 1.
- ✓ They are computationally efficient and simple to implement.

TABLE 3.3.
IMPLEMENTATION IN R OF B-SPLINES.

Implementation in R
Splines can be constructed in R using the package `splines` or `splines2` alternatively. The way B-Splines are implemented in R is slightly different than the way they are defined in the proposed literature [22]. Instead, a modification known as periodic is used, which is characterized by using equispaced and not-open knots: the endpoints of the interval are considered as knots as many times as the degree of the spline. This shapes the basis functions differently, but essentially their functionality is the same. A different basis means that we have a different coefficient for each basis function. This leads to having different components of the vector that represents the spline in the vector space (that is, a change of coordinates). This is no big deal as the main characteristics stay the same and both methods are equivalent.

Table 3.4 summarizes the main characteristics of both methods considered above.

TABLE 3.4.
COMPARISON OF B-SPLINES AND FOURIER SERIES.

Property	Fourier Series	B-Splines
Periodic Functions	Straightforward representation	Requires more complexity
Flexibility	Low	High
Derivative computation	Moderate	Very easy
Memory use of vector representation	High for some functions	Low

Note that, although the use of B-splines is convenient for any arbitrary function, other function approximations can also be used in different situations, as explained before.

For example, electromagnetic waves can be simulated using sinusoids, which are periodic. These sinusoids can be modulated to transmit information. There are several types of modulations depending on which attribute of the wave is used to transmit the data. Amplitude Shift Keying (ASK), Phase Shift Keying (PSK) and Frequency Shift Keying (FSK) are three of the most popular techniques.

In this context, one would prefer to represent the function by using Fourier coefficients, provided that Nyquist's theorem is satisfied [19], [23] (further information on the topic can be found in Oppenheim, Willsky and Nawab (2014) [21].

Hence, one must keep in mind that B-splines are not always the most optimal approximation method to use, and despite that this is the technique we consider all along the present work, our approach is applicable to any alternative curve fitting method.

Linear Least Squares Regression

If the function we want to represent is not a spline, we would find the spline that is *closest* to it in order to represent the function. The same thing applies to Fourier series analogously.

The way of defining which spline is closest is through least squares. The least squares method is proposed to find a solution to systems in which there are more equations than variables. These are called overdetermined systems [24]. This is our case because there are normally lots of input data points compared to the number of coefficients that we use to represent the function.

As we have a linear model made up of all the coefficients of the B-splines that represent our function, we can call this *linear least-squares regression*.

In least squares we find solution that satisfies that the sum of the squared errors is minimum. That is,

$$e_i = y_i - f(x_i \,;\, a_1, a_2, \ldots, a_n) \qquad \text{(Eq. 3.1)}$$

where:

- y_i are the observations, the measurements that we have obtained, that is, the original data.
- $f(.)$ is the target function, the one that we will obtain. It is the closest match to the data that we have observed.
- e_i is the error or difference between the observed data and the target function at the specific value of x. The vector of errors is represented by $e=(e_1, \ldots, e_N)$.
- x_i is such a specific value of x: the x axis value at which y_i was observed. For instance, it could be a time instant (e.g.: $x_i = 0.52ms$) or a position in space (e.g.: $x_i = 0.52mm$).
- a_i are the estimated coefficients, the parameters to be optimized by minimizing $||e||^2$. They are used to represent the function.

When we find $a_1, a_2, \ldots, a_n$ such that $||e||^2$ is minimum we have the best expression for $f(x_i)$ in the least squares sense. And once we have the function represented through the coefficients, we can use it for *interpolation*: following the reasoning behind functional data, this means that we can estimate samples at x positions (i.e. time instants) in which we had not observed the function originally.

TABLE 3.5.
IMPLEMENTATION IN R OF LINEAR LEAST SQUARES REGRESSION.

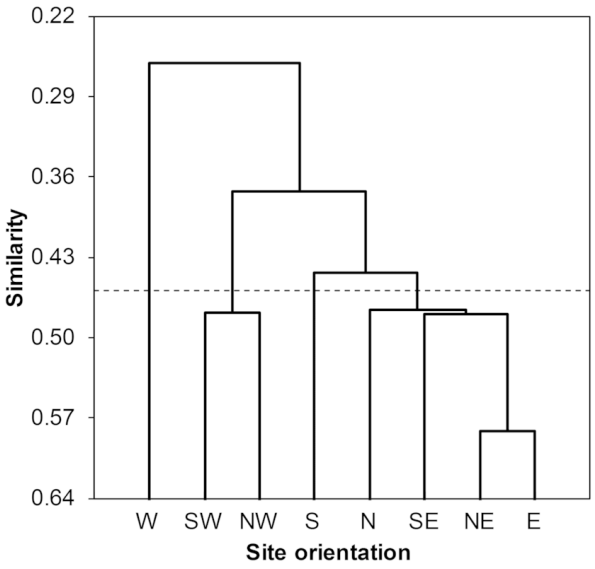
Implementation in R
Least squares can be implemented in R through the `lm(·)` function. It provides the linear least squares regression solution to the system passed as an argument. The system is formed, as shown in Eq. 3.1, by the input data, *y*, and the basis calculated using the `splines` package.

3.1.3 Clustering and Clustering Algorithms

Clustering is a form of unsupervised classification [25]. This means that input data will be grouped into a number *K* of classes, typically specified by the user, and that no *training* or *learning* data is used (i.e. we do not have any data to train the algorithm with).

For example, a high school teacher may want to group her students into different classes based on their academic performance. Information about their grades from primary school up to high school may be used. Say that she chooses to make three groups according to their level: low, medium and high. Using a clustering algorithm would make this grouping automatic for the teacher. Groups are also called *clusters* or *classes*.

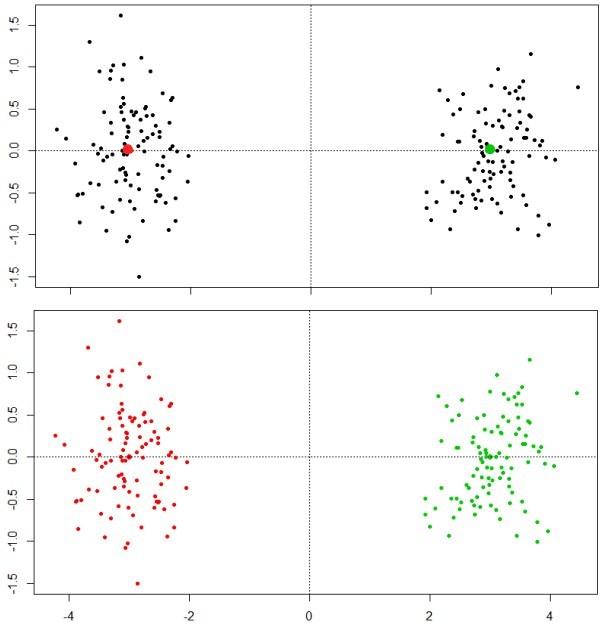
There is a massive amount of clustering techniques for multivariate data, but none has proven to work better than the others in every situation. Some of the most widely used types are discussed below.



- *Hierarchical clustering* is based on grouping data points as leaves of a tree, based on distance between these and the parent nodes. Points are grouped together if they are in proximity according to some distance criterium. As there are different ways of defining a "distance" between nodes, we find different algorithms according to those definitions. The hierarchical structure obtained can then be flattened by pruning the tree at certain heights.

Fig. 3.11. Hierarchical clustering of different orientations [26]. The *y*-axis measures proximity of data and clusters.

With respect to non-hierarchical clustering we have the following categorization:



- *Centroid based clustering* characterizes classes with a center. The groups, and consequently the centers, are chosen in order to "globally" minimize the distance of objects in a group to their respective center. Unlike hierarchical clustering, centroid based clustering calculates distances to "centers", which are not necessarily datapoints inside a cluster.

Fig. 3.12. Centroid based clustering. After selecting centroids for each cluster (top panel) each element is assigned to one of them (bottom panel).

- *Distribution-based clustering* makes the assumption that there is a certain underlying distribution model, such as a Gaussian distribution, which is used to find the clusters. Typically, the parameters of the model are optimized in each iteration of the clustering algorithm. A classic example of this is the expectation-maximization (EM) algorithm [28], as shown in Fig. 3.13.

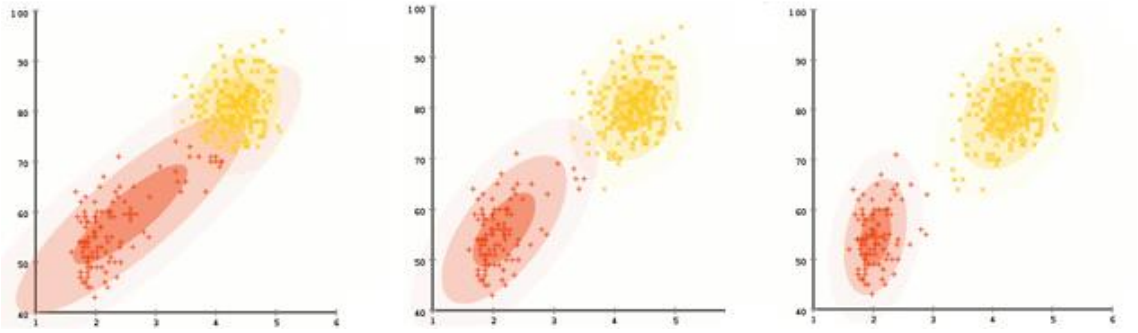


Fig. 3.13. EM algorithm with a mixture of Gaussians model [27]. The mean and standard deviation of the 2-dimensional Gaussians are adjusted in each iteration from left to right according to the expectation-maximization algorithm.

- *Density-based clustering* is similar to hierarchical clustering as the grouping is done according to distance between datapoints. The difference between them is that, in density-based clustering, points that are further away are labeled as noise, whereas in hierarchical clustering new clusters would be created for those outliers.

How do we know which algorithm to use? The answer to this question is not simple; researchers tend to select a technique that has proven to provide good results for the kind of data under analysis, or at least for a wide collection of situations. In particular, we consider here the K-Means algorithm, probably the most popular clustering technique in place in numerous research fields.

K-Means and K-Medoids

Two important reasons that make K-Means so widespread are that it provides better results than other clustering techniques with little information and that it is simple to implement.

Fig. 3.14 and Fig. 3.15 illustrate how the algorithm works. Observed datapoints are plotted as black dots, whereas cluster centers or *centroids* are colored and circled.

Step 1: Initialize the algorithm with "K" *random* points or elements from the dataset. These will be the initial centroids. In this case we have K = 3.

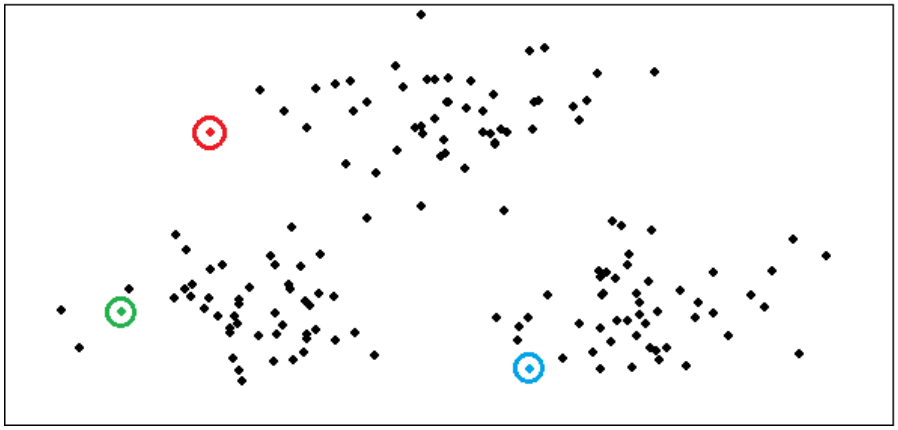


Fig. 3.14. Random initialization of K-Means.

Step 2: Assign each data element to exactly one cluster (represented by a centroid). The data is grouped according to the distance to each of the possible cluster centers (green, red and blue) [29]. This is shown in Fig. 3.15 (top panel).

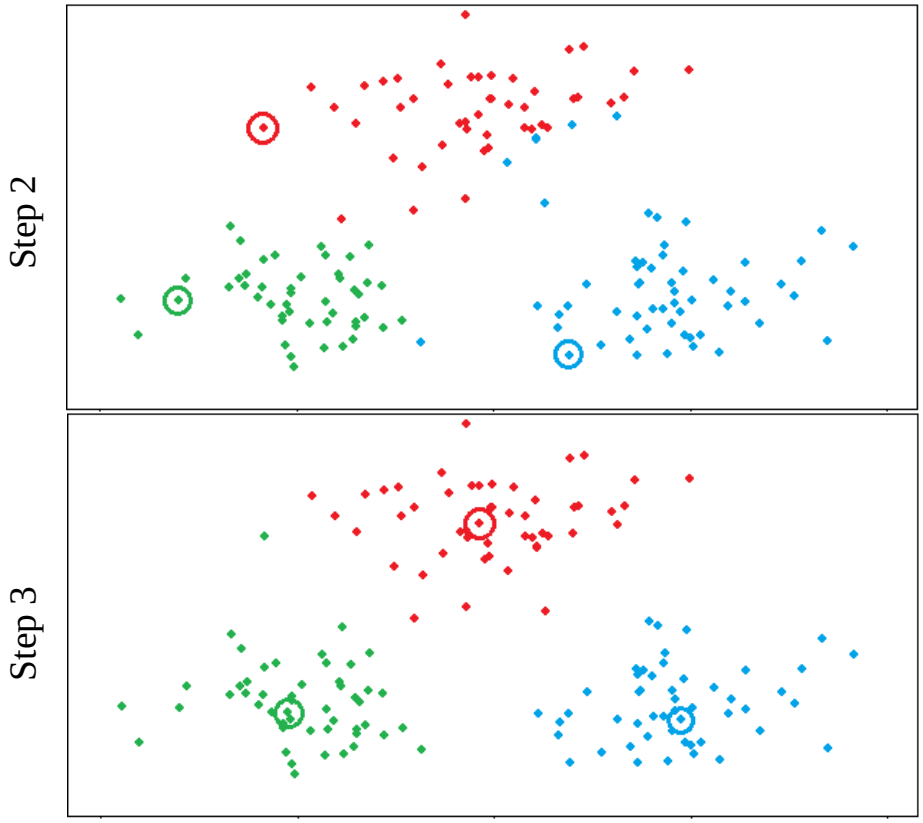


Fig. 3.15. K-Means convergence.

Step 3: Update the centroids of each cluster by calculating the component-wise mean of all the datapoints' coordinates (hence the name K-*Means*), as shown in Fig. 3.15 (bottom panel).

Now, we repeat steps 2 and 3, reassigning the datapoints to new clusters with the new centroids and recomputing the centroids for each new cluster after that. By repeating this, the algorithm eventually converges (i.e. does not produce a different clustering output), or reaches a maximum number of iterations.

K-means is also known to yield a minimum[2] of the so-called *distortion*:

$$D = \sum_{x \in X} ||x - c||^2,$$

where X is the dataset, and c is the nearest center to the datapoint x. More details on distortion will be discussed in subsection 3.2.3 (*Clustering Evaluation Techniques*).

All in all, we see that the centroids move around and end up in a suitable, center-representative, position. Additionally, it can be seen from the way in which they are computed that the centroids are not necessarily datapoints.

In contrast, an alternative algorithm known as *K-Medoids* [30] is based on the same principle as *K-Means* but choosing as a centroid a point that belongs to the cluster. The most popular variation of this method is the Partitioning Around Medoid (PAM) [31] algorithm, which eliminates the randomness when selecting the centroids [32].

[2] It is not guaranteed whether it is a local or global minimum of the distortion.

It can be anticipated that there must be a more convenient way of initializing the K-Means algorithm than the random initialization introduced in Step 1. K-Means initialization is a whole new world by itself, and several techniques have been proposed up to date to address this problem. For a review, refer to Celebi, Kingravi, and Vela (2012) [10].

K-means++ is probably the most common option; it chooses a random initial point, but the choice of the second point is conditioned on the choice of the first one. In other words, there is a higher probability of choosing a point that is further away from the first one chosen. Points that are selected further away tend to provide better convergence and more accurate results.

TABLE 3.6.
IMPLEMENTATION IN R OF K-MEANS, K-MEDOIDS AND K-MEANS++.

Implementation in R
K-means is implemented in R and can be accessed by calling the `kmeans()` function. The most common variation of K-Medoids is the PAM algorithm, implemented in R via the `pam()` function. An implementation of K-Means++ is found in the `LICORS` package by calling `kmeanspp()`. Additionally, other implementations can be found online but this package has been chosen due to its simplicity and good performance according to the original definition of K-Means++ by Arthur and Vassilvitskii (2006) [33].

The main goal of this project is to propose and test a new algorithm to initialize *K-Means*, using another principle to provide the initial centroids. This will be discussed in subsection 3.2.2 (*MBD as a Solution to K-Means Initialization Problem*).

3.1.4 Clustering Evaluation Techniques

How do we determine if a clustering result is better than another one? To quantify this idea of "being better" we must find a technical way for comparison.

If we input some data to any non-hierarchical clustering algorithm, a *label* vector that determines which data belong to which cluster is produced. A simple representation of this is presented in Fig. 3.16.

Datapoint (Input)	Cluster Centroid	Label (Output)
(-3,0)	(-3,0)	1
(-4,0)	(-3,0)	1
(2,0)	(2.5,1)	2
(3,2)	(2.5,1)	2
(-2,0)	(-3,0)	1

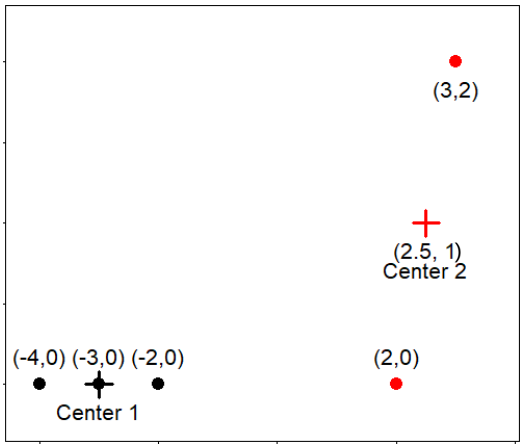


Fig. 3.16. Label vector as output of clustering (left). 2D-graphical visualization of labels as colors (right).

In this case we have considered K = 2 (number of clusters to be extracted), and input datapoints of dimension D = 2.

It can be seen that the input data can have any number of dimensions and that in the case of functional data, it becomes infinite. As explained in footnote 1 in section 3.1.1 (*Multivariate and Functional Data*), functional data can be discretized to have a number of dimensions equal to the number of observations of the function that we have fitted.

A more realistic example and of more use to the project would be taking two functions (i.e. K = 2) observed at the same time instants: a sine and a cosine, with some level of noise. Say we have observations of these sinusoids every 0.1 seconds. See Fig. 3.17.

Datapoint (Input)	Label (Output)
(-0.188,0.155,-0.052, 0.774, …)	1
(0.013, -0.173, 0.246, 0.991, …)	1
(0.476, 0, -0.487, 1.045 …)	1
(0.885, 1.49, 1.433, 0.98, …)	2
(0.875, 0.882, 0.870, 0.867, …)	2
(1.48, 0.583, 0.905, 1.303, …)	2

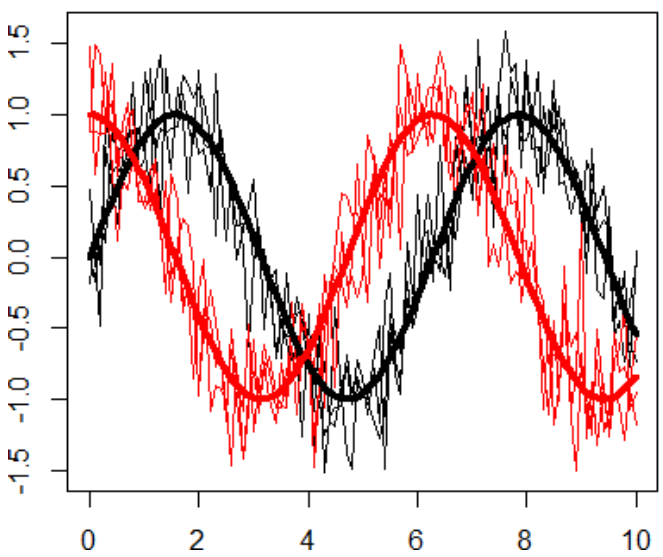


Fig. 3.17. N-dimensional vectors of observations from continuous functions (left panel). Correct assignment of clusters, coded by color (right panel).

The cases depicted in Fig. 3.16 and Fig. 3.17 are examples of situations in which we could analyze the clustering output produced. Clustering evaluation techniques can be classified [34] into:

- *Internal* techniques, which are those that use the input data to evaluate clusters, and
- *External* techniques, which are the ones that use the output labels to evaluate clusters. In general, they compare two sets of (output) labels to assess the agreement between them.

Nevertheless, there is an important distinction to make between the *computed* labels and the *real* or *original* labels. In other words, there is a crucial difference between the groups that the algorithm makes and the real groups that the data come from.

For instance, taking the same example of sine and a cosine signals measured in the presence of noise proposed in Fig. 3.17, if the noise is sufficiently high, the distinction of the sine and the cosine is made harder and can mean that the classification is not done correctly. That is, what came from a sine signal is classified as a cosine, and vice versa.

Typically, the true labels are not known in the clustering problem, but simulated data or certain *a priori* knowledge about real data can provide such labels. Accordingly, external cluster evaluation techniques are the ones that can be used to compare original labels with computed labels to determine clustering precision.

Additionally, one cluster alone can be evaluated using internal measures, but not external measures. Hence, both external and internal measures are needed to have an overall idea of the performance of a clustering method.

Once this difference is made clear, we can use the subsequent, commonly used, measures shown in Table 3.7 to evaluate the performance of our initialization method.

TABLE 3.7.
CLUSTERING EVALUATION TECHNIQUES.

Name	Description	Range	Type
Purity	Each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned datapoints (with respect to the most frequent class in the cluster) and dividing by N [34].	[0, 1] (real)	External
Rand index	Defined as the number of pairs of objects that are either in the same group or in different groups in both label sets divided by the total number of pairs of objects. When two label sets agree perfectly, the Rand index has a value of 1 [35].	[0, 1] (real)	External
Adjusted Rand Index (ARI)	A problem with the Rand Index is that the expected value of the Rand index between two random partitions is not a constant. This problem is corrected by the Adjusted Rand Index that has an expected value of 0 in the case of random clusters [35].	[0, 1] (real)	External
Distortion	For each cluster, the squared distances of each datapoint to the corresponding center are summed.	$[0, \infty)$ (real)	Internal
Iterations	The number of iterations until the K-Means clustering algorithm converges, once it has been initialized.	{1,2,3 ... } (natural)	Other
Execution time	The time it takes for the clustering algorithm to compute the output labels. It clearly depends on the specific implementation of the methods.	$[0, \infty)$ (real)	Other

Occasionally, these measures will not be enough to determine which clustering algorithm works better. We have defined a *correctness*[3] index that returns the number of correctly

[3] Computing the probability of error, P, is the same as computing the correctness, C. They are related through C = 1 – P.

classified datapoints as a percentage. To do so, all permutations of the original labels are computed, and the maximum percentage of correctly classified points of all permutations yields the correctness[4]. It is an external measure.

However, correctness is not used as frequently as purity or ARI because of the time it takes to be computed, especially if the number of clusters is large.

Note that whereas purity, ARI and correctness are bounded between 0 and 1, distortion, iterations and execution time can take any positive value.

Finally, it can be said that the higher the value (i.e. the closer to 1) of purity, ARI and correctness, and the lower the value (i.e. the closer to 0) of the distortion, the iterations and the execution time, the better the clustering algorithm will be.

3.1.5 Bootstrapping

Bootstrapping, which has been mentioned before but not defined, is a technique based on estimating the sample distribution of a statistic by resampling from the available data [36] which are assumed to be samples from a certain population. The underlying idea is to consider these samples as a new population and to sample it again, in order to obtain the so-called bootstrap replicas [37], as pictured in Fig. 3.18.

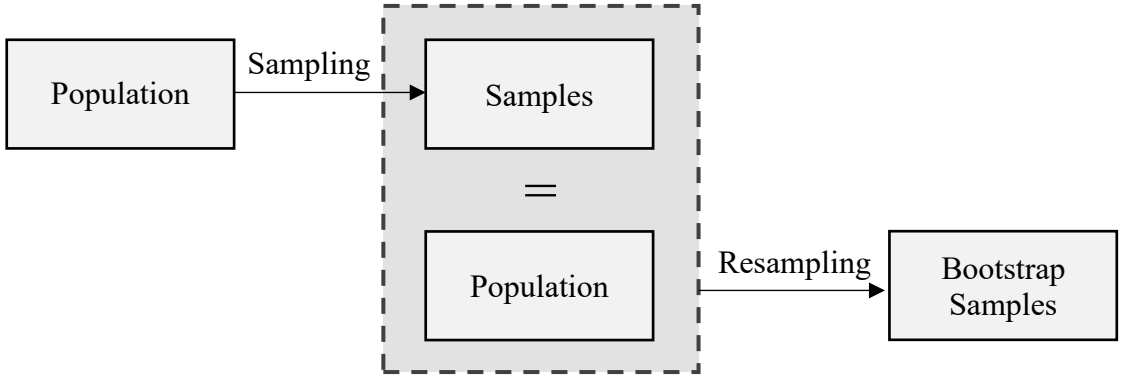


Fig. 3.18. Bootstrapping procedure.

The sample size of the bootstrap replicas is the same as the initial sample size[5]. For example, if we have 5 input samples, we would have bootstrap replicas with 5 elements, which are drawn *with replacement*, leading to not necessarily identical sets. This is illustrated in the diagram in Fig. 3.19.

It can be shown that the mean of a statistic (i.e. adding all statistics and dividing by B, as depicted in Fig. 3.20) of the bootstrap replicas is a more accurate representation of the statistic of the actual population than that provided by the initial sample, which can be seen as a loose application of the law of large numbers [38].

Some examples of the statistics that can be calculated are standard deviations, means, covariances, and others that will be addressed in section 3.2 (*Techniques Used*).

---

[4] Another measure that is commonly used in statistics is the *confusion matrix*. It provides the correctness value as well as which clusters are confused with which. For example, sines can be confused with cosines, exponentials with x^2, and so on. This measure will not be used in the project because of its computational complexity and because it provides additional information that is not relevant to the goals of the project.

[5] In fact, as long as the bootstrap sample size is big, the same theory holds.

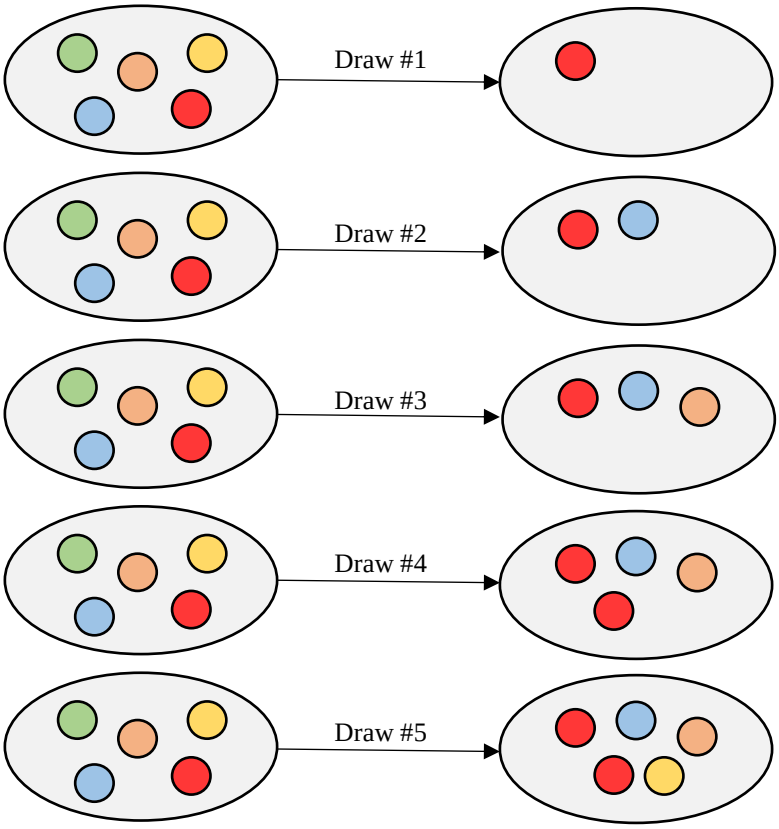


Fig. 3.19. Resampling with replacement. The sample size is the same after Draw #5, but the elements inside are different. The computation of some statistics for both datasets would result into slightly different values.

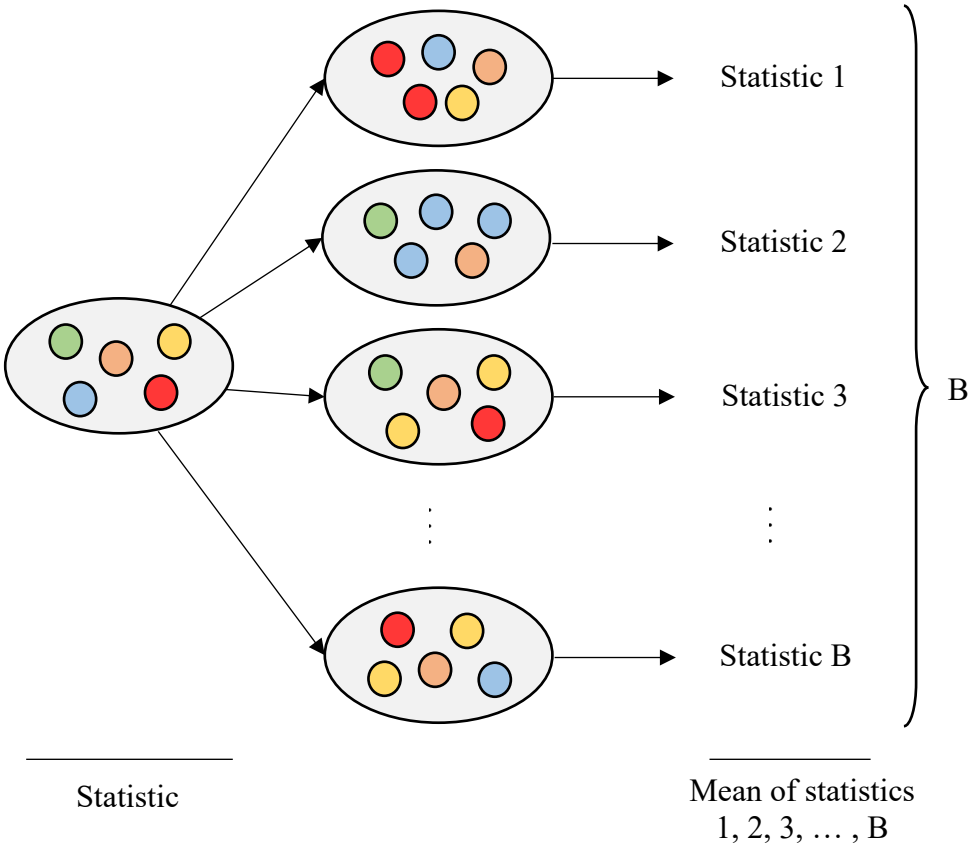


Fig. 3.20. Repeating resampling. This yields B other samples, generated from the initial one on the left. In a way, this is like generating a larger sample from the original population with the knowledge of only a few samples.

TABLE 3.8.
IMPLEMENTATION IN R OF BOOTSTRAPPING.

Implementation in R
A bootstrap implementation is found in the `boot` package in R. The `boot()` function requires specifying the initial data matrix, the statistic to be bootstrapped and the number of bootstrap replicas to generate the bootstrapped statistic of interest [39].

3.1.6 Modified Band Depth (MBD)

The notions of depth and band depth will be introduced in this subsection as a foundation for MBD, a concept that will also be detailed in this light.

Depth

The term *depth* in multivariate statistics refers to the degree of centrality of a datapoint, as a generalization of the univariate median. Accordingly, for a univariate dataset, the deepest point is a measure of the central tendency of the dataset. As we get further and further away from the central point, depth decreases [40].

This idea can be extended to functional data, in which the deepest curve can be informally thought of as "the one whose graph is located in the middle". In Fig. 3.21 the central (i.e. deepest) curve is the black one.

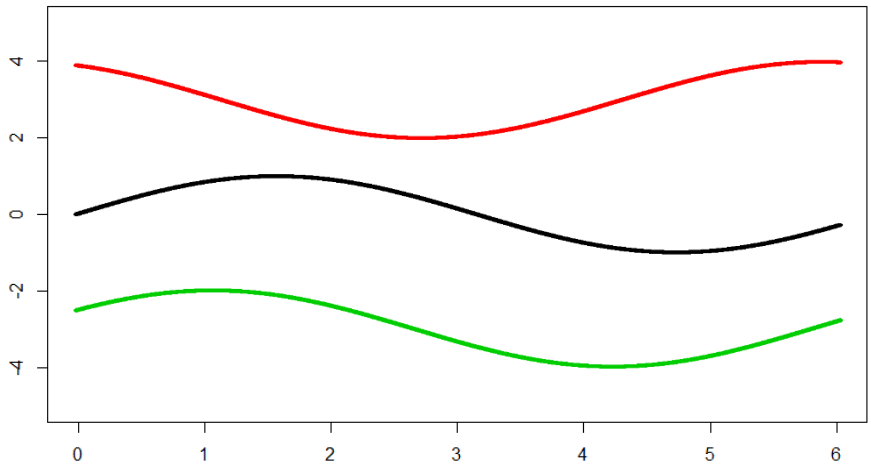


Fig. 3.21. Depth notion for functional data. The black curve is the deepest one.

Unfortunately, as it happens frequently in statistics, the formal definition is not as simple as the non-technical one. Whereas it is easy to choose a central curve by inspection in these simple scenarios, we need a more formal definition for the more complicated cases.

There are several ways of defining a suitable notion of depth as described in Cascos, López and Romo (2011) [40], where more specific details are given about depth functions. It is an outstanding reference for obtaining a deeper insight in the matter.

For the context of this project, we briefly describe the Band Depth (BD) and the Modified Band Depth (MBD), introduced by López-Pintado and Romo (2009) [12]. MBD has become a very popular depth measure due to its low computational cost, in contrast to most alternatives found in the literature.

Band Depth (BD)

BD is defined for functions or multivariate data represented in parallel coordinates [41].

Fig. 3.22 shows the sampled, finite-dimensional, version of Fig. 3.21. It can be seen that the red and the green dotted curves surround the black dotted curve at all *x* instants. This is referred to as the black curve (multivariate datapoint) being inside the band defined by the red and green curve (multivariate datapoints).

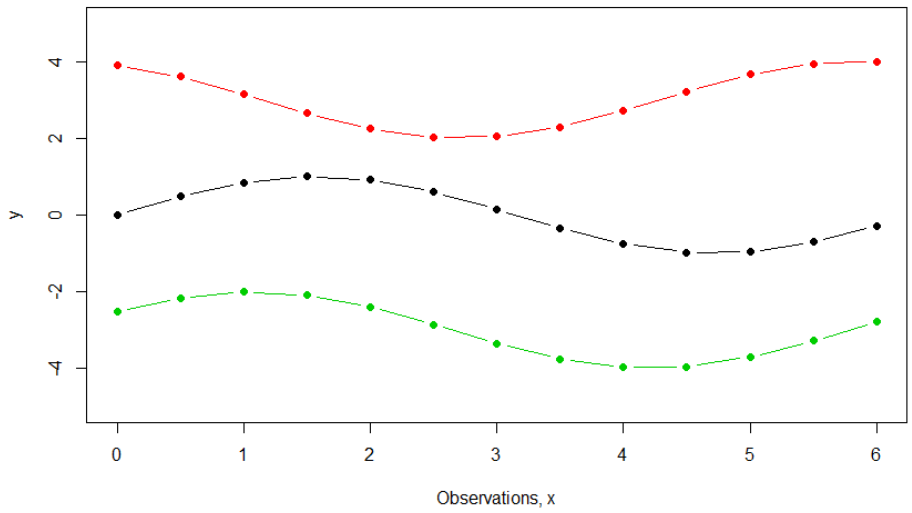


Fig. 3.22. Band depth notion. The black curve is inside the band defined by the red and green curves.

In this definition of depth, centrality is based on the number of bands that contain a certain function or datapoint. The more bands enclosing a datapoint, the more central it will be.

The finite-dimensional version of a band defined by two or more curves $x_1, \dots, x_j, j \geq 2$, [40] is the set:

$$B(x_1, \dots, x_j) = \left\{ y \in \mathbb{R}^d : \min_{1 \leq i \leq j} x_i^{(k)} \leq y^{(k)} \leq \max_{1 \leq i \leq j} x_i^{(k)}, \ 1 \leq k \leq d \right\},$$

where $x_i^{(k)}$ and $y^{(k)}$ are the *k-th* components of x_i and y respectively. For a point *x*, we define

$$BD^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \cdots \leq i_j \leq n} \mathrm{I}\{x \in B(x_{i_1}, x_{i_2}, \dots, x_{i_j})\}, \qquad \text{(Eq. 3.2)}$$

where *n* is the number of datapoints in the dataset, x_i, and I{·} is the indicator function. $BD^{(j)}(x)$ expresses the proportion of bands $B(x_{i1}, x_{i2}, \dots, x_{ij})$ determined by *j* different curves $x_{i1}, x_{i2}, \dots, x_{ij}$ containing the whole graph of *x*.

For instance, say that we have $j = 2$ curves to define a band. To calculate the $BD^{(2)}(x_i)$ of a datapoint x_i, we have to consider all pairs of curves (i.e. x_1 and x_2, x_2 and x_3, x_1 and x_3, etc.) and check whether x_i is *entirely* inside the band or not. In case it is, then the indicator function will return 1; otherwise, it will return 0. By counting all the cases in which the datapoint is embedded in a band and normalizing by $\binom{n}{j}^{-1}$, we get the band depth for that point.

An example of two datapoints being inside and outside a band is depicted in Fig. 3.23: y_1 is inside the band defined by x_1 and x_2, and y_2 is outside it most of the time. The contribution to the band depth given by the band defined by x_1 and x_2 is 1 for y_1, and 0 for y_2.

Fig. 3.24 shows the band defined by more than two curves. It is the region of the plane determined by the most external curves at every coordinate/time instant.

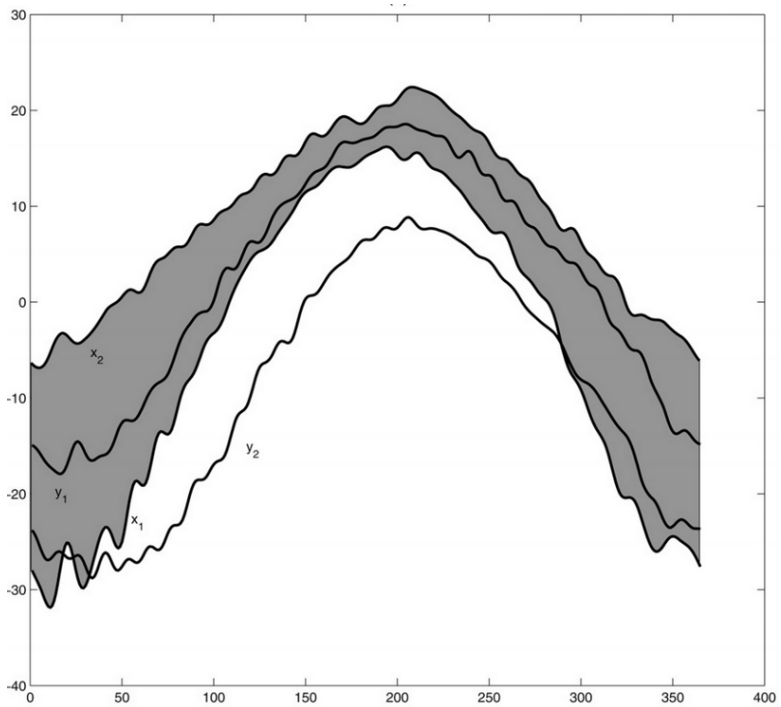


Fig. 3.23. Band inclusion and exclusion [12]. y_1 is entirely inside the band defined by x_1 and x_2, whereas y_2 is not.

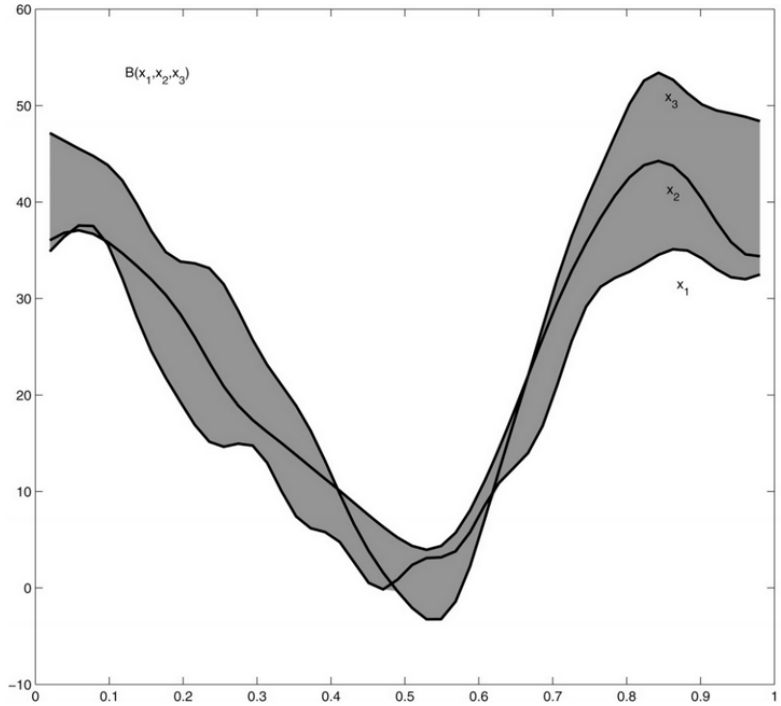


Fig. 3.24. Band defined by three curves as the gray region [12].

Finally, to define the BD, we choose the number of curves $J \geq 2$ that will define the bands and calculate:

$$BD_J(x) = \sum_{j=2}^{J} BD^{(j)}(x). \qquad \text{(Eq. 3.3)}$$

Hence, the value of BD_4 is an addition of the values $BD^{(4)} + BD^{(3)} + BD^{(2)}$. A J value of 2 or 3 is usually enough for all applications.

Band depth is used because it has a low computation complexity for high dimensional datasets, but has the disadvantage of allowing multiple ties in the depth values.

Modified Band Depth (MBD)

MBD is a less restrictive version of BD as the depth of a curve does not depend on the whole curve being inside a certain band. Instead, the finite dimensional version of MBD is based on computing the mean number of coordinates that are inside a band; d is the number of coordinates, and $x^{(k)}$ the *k-th* components of datapoint x.

$$MBD^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \cdots \leq i_j \leq n} \frac{1}{d} \sum_{k} \mathrm{I}\{x^{(k)} \in B(x_{i_1}, x_{i_2}, \ldots, x_{i_j})\} \qquad \text{(Eq. 3.4)}$$

The main advantage of MBD over band depth is that it is less computationally intensive.

3.2 Techniques Used

Now that the essential theoretical background has been summarized, a closer look at how all this theory helps to improve functional data clustering must be taken. This subsection corresponds to tasks D1.1 and D1.2 of Table 2.1 (*Task Breakdown and Duration*).

From the perspective of the project as a case study, it is important to determine the techniques that will be used to obtain results and reach meaningful conclusions. It is crucial to perform both a quantitative and a qualitative analysis of the method.

A comparison between the different design techniques that have been mentioned before will be discussed, as well as the reasoning behind the choice of the final implementation of the solution.

3.2.1 B-Splines for Function Approximation

In the literature review section two alternatives for function approximation have been presented: Fourier series and B-Splines. Fourier series provide a useful representation for periodic functions, but are more computationally intensive to implement than B-Splines.

From a qualitative perspective, B-Splines are more convenient than Fourier series for the following reasons, as described in Table 3.4 of section 3.1.2.

- ✓ *Simple vector representation of a function and its derivatives*. It is very easy to differentiate splines as it only involves computations on the basis. This provides fast access to additional information for clustering.
- ✓ *Fixed finite basis for some given knots and order*. If all input data has the same length, having a fixed basis reduces the computational complexity of the basis calculations.

- ✓ *Computational efficiency and simple implementation*. There is an R package that implements B-splines efficiently and it is easy to use once the input data has been collected into a matrix.

Drawing all these points together, and for the reasons explained above, the function approximation technique used in this work is B-splines. Fig. 3.25 summarizes the process.

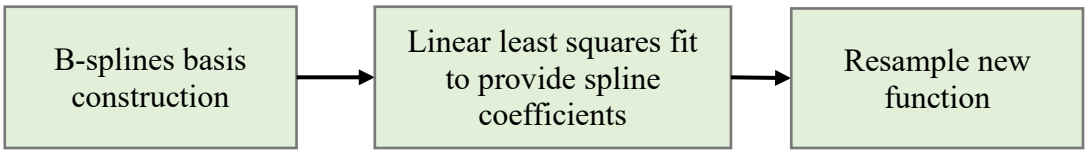


Fig. 3.25. Steps in B-spline approximation.

3.2.2 MBD as a Solution to K-Means Initialization Problem

As previously mentioned, K-Means is a clustering algorithm that relies on random initialization of centroids. A correct initialization of K-Means would improve the robustness and overall clustering results.

MBD is used here to derive a solution to this initialization problem. The best way to picture how we can benefit from MBD for functional data clustering is through the information flow diagram from input to output shown in Fig. 3.26.

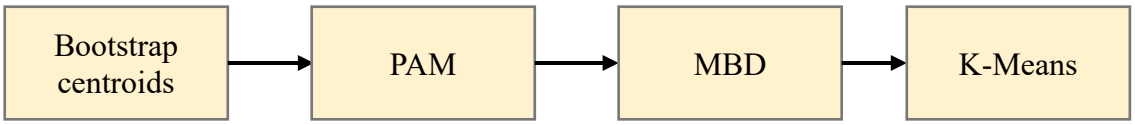


Fig. 3.26. MBD as a Solution to K-Means initialization problem.

The input to this system is the output of Fig. 3.25, the new samples of the approximated function. These new samples are bootstrapped B times. For each of the B bootstrap samples, we run pure K-Means (with random initialization) to obtain some centroids.

These centroids are collected into one dataset. In total we have $K \cdot B$ centroids, where K is the number of clusters (i.e. we have K cluster centroids for each of the B replicas). Now we have a collection of points in space that correspond to the centers of the different clusters of the bootstrap replicas. These elements are estimators of the real cluster centers; the variability present in this collection of centroids comes from both the bootstrap step and the random initialization of K-means. They are expected to form tighter groups than the original dataset, hence being easier to cluster.

At this stage, we form groups of bootstrap centroids by using Partitioning Around Medoids (PAM) in order to reduce randomness and to get a more robust output, which does not depend on K-means initialization.

Finally, MBD is applied to find the deepest point inside each cluster formed by PAM. The deepest points are chosen to initialize K-Means. This is seen in Fig. 3.27.

Following this procedure, K initial points will be found, and clustering algorithms would have been run a total of $B+2$ times, counting PAM and *K-Means*' final execution after initialization.

This part of the method is the one being tested. The advantages and disadvantages of choosing MBD-based initialization of K-Means will be detailed in section 5 (*Conclusions*).

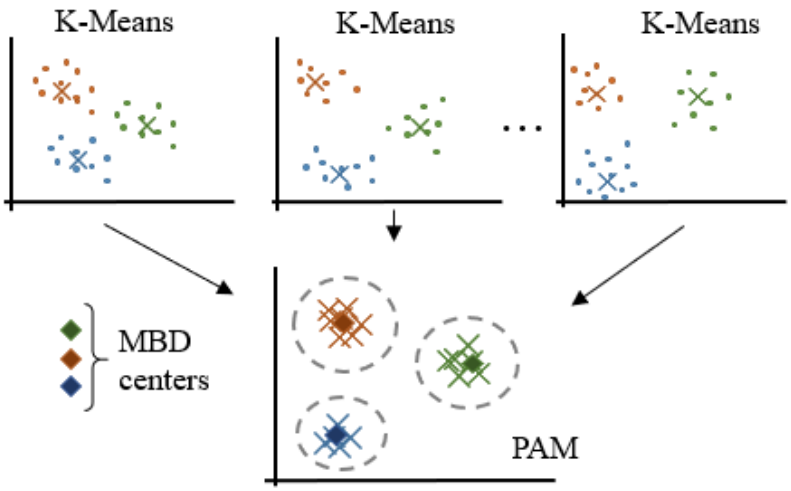


Fig. 3.27. K-Means initialization diagram.

3.2.3 Clustering Evaluation Techniques Used

The quantitative measures used to evaluate clustering results are the following:

1. Correctness
2. Purity
3. ARI
4. Distortion
5. Iterations
6. Time (Execution)

A description of these can be found in section 3.1.4. This is what we will call the CoPADIT measures, an acronym formed by the first letter or letters of the names of the measures to be used.

- Correctness, purity and ARI are referred to as *accuracy measures*.
- Iterations is a *convergence measure*, that indicates how rapidly the K-Means algorithm reaches a stable set of centroids.
- Cluster *dispersion measures*, like distortion, assess the spread of the datapoints that belong to a cluster.
- A method's execution time evaluates its *computational cost*.

In order to qualitatively check if the proposed method is worth using, the clustering result we obtain is to be compared with the ones yielded by other existing initialization methods.

There are five methods to be compared, mentioned in the *List of Acronyms* in the introductory pages, and defined as:

1. *KM*: K-Means with random initialization.
2. *MVMBD*: K-Means initialized with MBD for multivariate data.
3. *FMBD*: K-Means initialized with MBD after B-splines function approximation and resampling.
4. *KMPP*: K-Means++.
5. *FKMPP*: K-Means++ after B-splines function approximation and resampling.

We will refer as *three-way comparison* to comparing methods 1, 2 and 4, and as *five-way comparison* to comparing all five methods with one another.

3.2.4 Summary

On the whole, a general picture of the method is presented in Fig. 3.28, condensing into a flow diagram the techniques described in this section (*Methods*).

Table 3.9 recapitulates all the parameters that can be tuned by the user in the FMBD method as explained throughout this chapter.

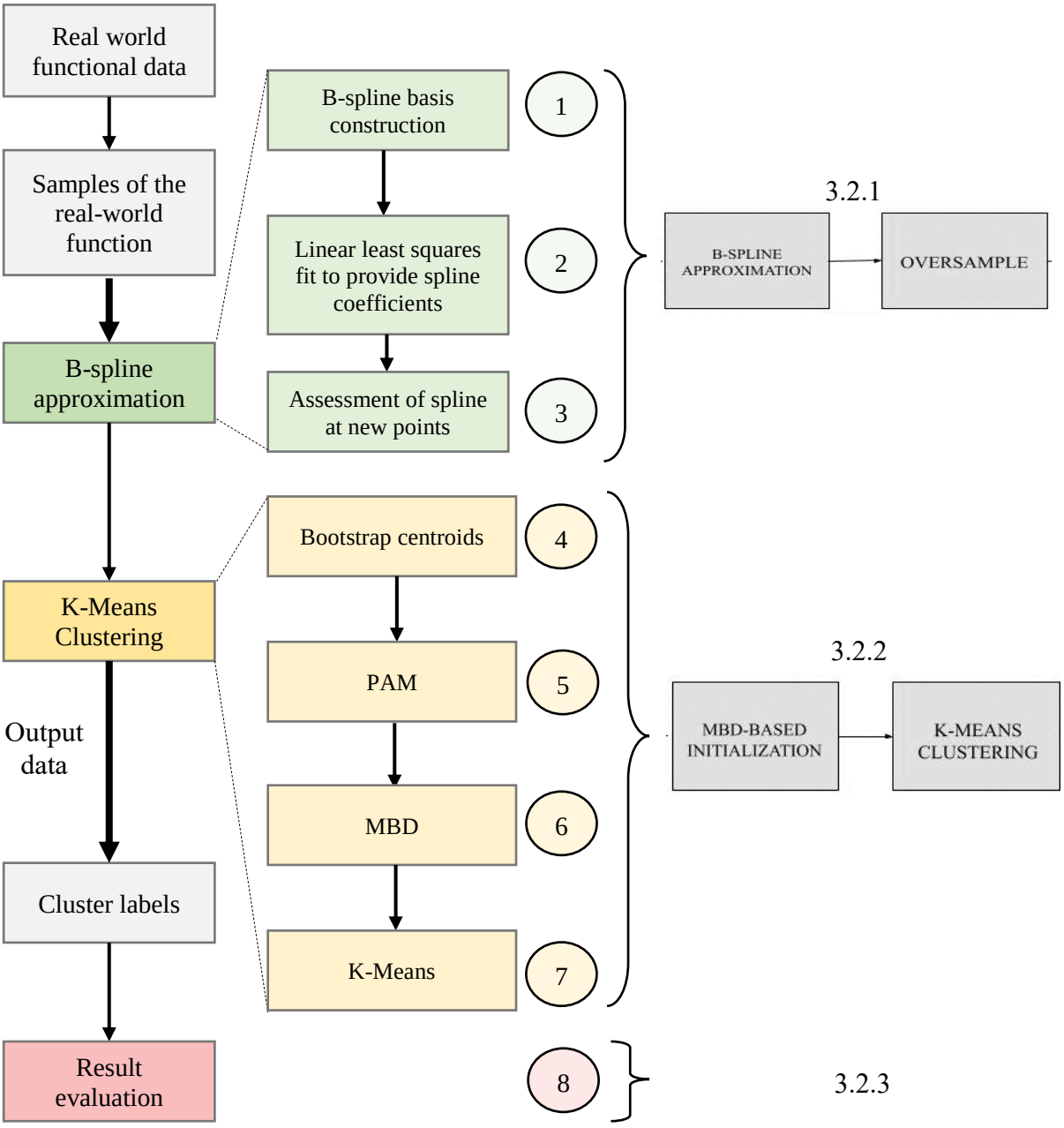


Fig. 3.28. Overall picture of the method.

TABLE 3.9.
PARAMETERS DEFINED BY THE USER IN THE FMBD METHOD.

Parameter	Step of the overall picture of the method	Use
Intercept	1	Determines if the function has a *y*-intercept (true) or starts at zero (false).
Degree of polynomial (DP)	1	Defines the degree of the polynomial used to compute B-splines.
Degrees of freedom (DF) / knots	1	Indicates where the knots will be located prior to estimating the function.
Oversampling Factor (OSF) / Observation x-axis (vect)	3	Determines where we want to observe the approximated function.
Bootstrapping replicas (B)	4	Sets the number of times bootstrapping is repeated

3.3 Models for Testing

We test our method (FMBD) and compare it to alternative initialization algorithms on different models to represent situations characterized by some parameters. In particular, the most relevant ones are:

- *Number of clusters*. Each cluster represents one particular function. For example, we may have a "sine" cluster, an "x^2" cluster, and so on.
- *Number of datapoints per cluster*. In our experiments, all clusters will be of the same size.
- *Length of each datapoint*, or equivalently, the number of observations of each function. This is set by the *x* vector which represents the horizontal axis. It can also be understood as a time vector.
- *Noise level*. Additive White Gaussian Noise (AWGN) is the most common type of noise in signal processing and communications. It mimics the effect of measurement errors, superposition of several signals at the point of measurement, and the naturally-occurring thermal noise [42].
  - It is *additive* because the noisy signal is the sum of the clean signal and the noise.
  - It is *white* because it affects all frequencies by the same amount (i.e. the spectral power density is flat).
  - It is *Gaussian* because it follows a Gaussian distribution of mean $\mu = 0$ and standard deviation σ.

Additionally, noise is used to add intra-cluster variety. The generation of two datapoints from the same original function leads to two identical points. In the presence of noise, these will be separated from the theoretical center represented by the original function.

A 2-dimensional plot of a cluster center and the datapoints of each cluster in the presence of AWGN is presented in Fig. 3.29.

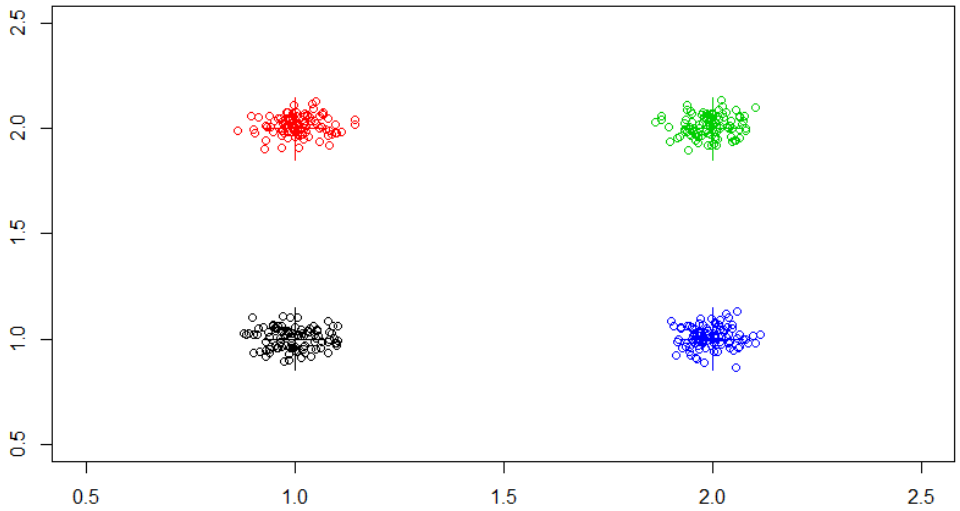


Fig. 3.29. Noisy 2D representation of clusters. The datapoints belonging to each cluster are piled around their centers.

Using these same centroids, more noise can be added by increasing the value of the AWGN's standard deviation. This is shown in Fig. 3.30.

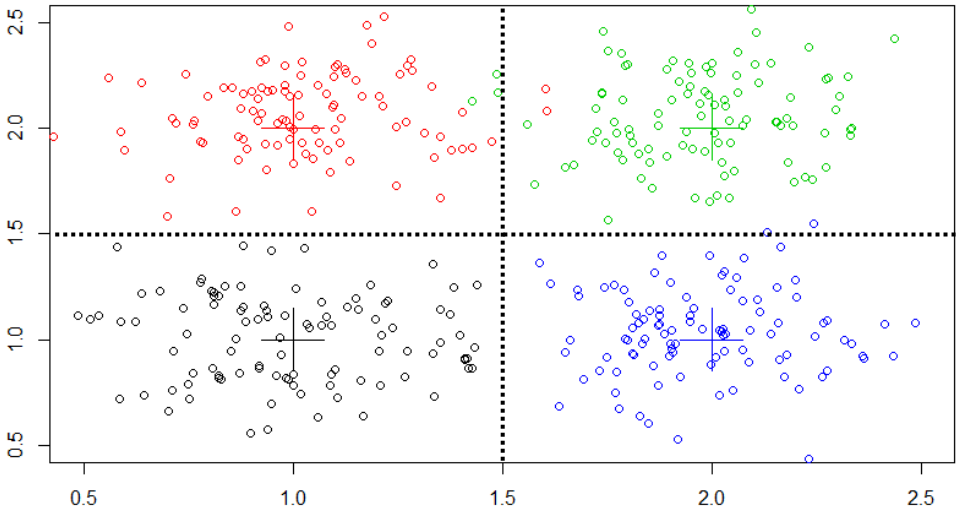


Fig. 3.30. High noise scenario.

Analyzing the figure, it is clear that the higher the noise, the harder it is to distinguish the original cluster. In particular, K-Means performs groupings according to distance which implies that even if the original cluster centers are used, some points will be wrongly classified as they are closer to an incorrect cluster centroid than to the correct one. This is the case, for example, of three of the green points represented in Fig. 3.30, which are in the Voronoi region [43] of the red cluster. In this case, these would be classified as red points and not green ones.

All in all, noise is presented as a confusion element. The more noise there is, the harder it is to provide a satisfactory result. The standard deviation of the noise is the parameter to be tuned in order to evaluate the different clustering methods in sub-optimal situations.

We illustrate the role of each parameter described before in the following four models.

Model 1

The model is presented in Fig. 3.31. Each function graphed in the figure gives origin to a cluster. In this case we have the set of generating functions described in Table 3.10.

TABLE 3.10.
MODEL 1 FUNCTION DEFINITION.

Cluster Number	Function
1	$y = x$
2	$y = 2 \cdot (x - 0.5)^2 - 0.25$
3	$y = -2 \cdot (x - 0.5)^2 + 0.3$
4	$y = 0.6 \cdot sin(2\pi \cdot x)$

Under Gaussian noise, the seed of each cluster is given by the noiseless generating function that created it[6].

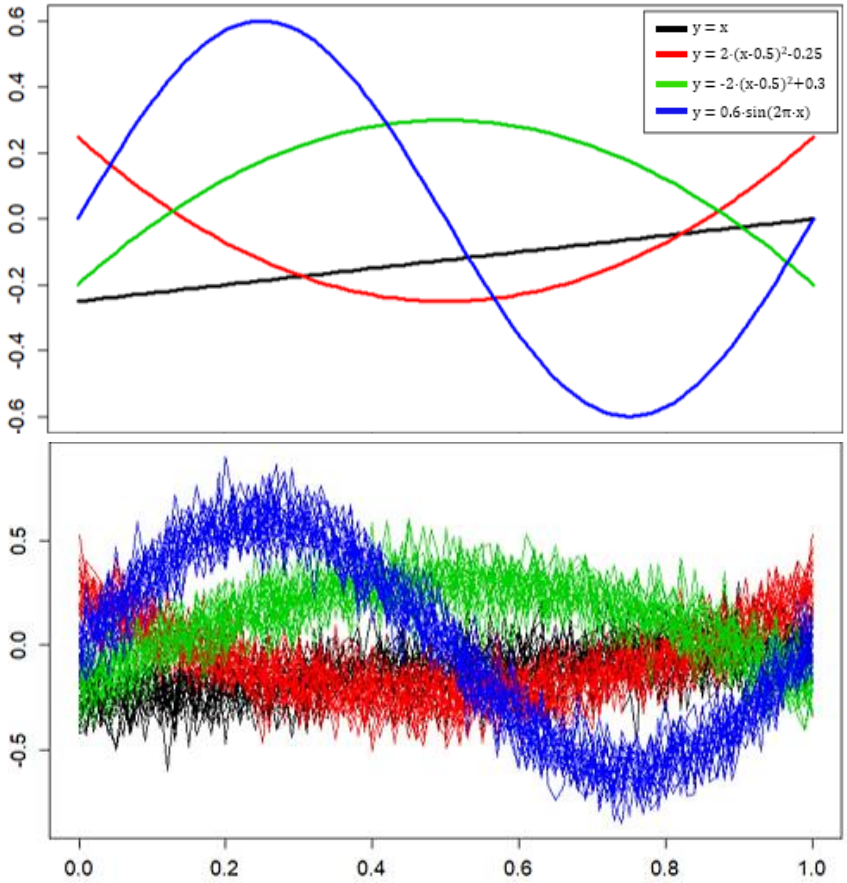


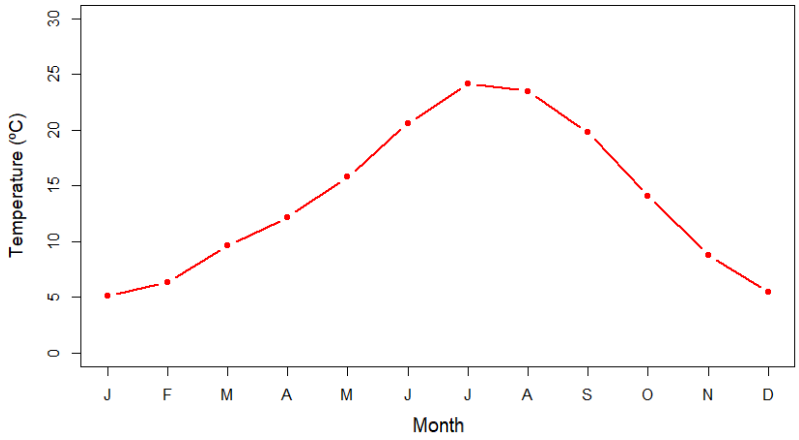
Fig. 3.31. Model 1, noiseless (top) and noisy with sigma = 0.1 (bottom), observed at 101 equispaced time intervals.

The *x*-axis is defined from 0 to 1, with a step of 0.01. There is a total of 101 observations, including the one at $x = 0$. To represent this in a euclidian space, 101 dimensions would be required to illustrate one point.

Model 1 was conceived to somewhat resemble the monthly average temperatures in a year for different climates.

[6] This is used in digital communications, known as a Gaussian channel [43]. The centers correspond to the symbols transmitted. The symbols received are clustered according to the Voronoi regions.

In the case of the city of Leganés in Spain, we can picture the temperature curve to be a convex quadratic function.



	J	F	M	A	M	J	J	A	S	O	N	D
Mean Temperature (°C)	5.1	6.3	9.6	12.2	15.8	20.6	24.2	23.5	19.8	14.1	8.8	5.5

Fig. 3.32. Mean temperature in Leganés [33].

Model 2

Model 2 is a modification of model 1, having a sinusoid with higher frequency. The purpose of this model is to assess how rapidly-changing signals condition the clustering results. There are also four functions that define each cluster in this model as described in Table 3.11.

TABLE 3.11.
MODEL 2 FUNCTION DEFINITION.

Cluster Number	Function	Color
1	$y = x\text{-}0.5$	Black
2	$y = (x\text{-}0.5)^2\text{-}0.8$	Red
3	$y = -(x\text{-}0.5)^2+0.7$	Green
4	$y = 0.75 \cdot \sin(8\pi \cdot x)$	Blue

Apart from the frequency increase in cluster number 4, some scaling and vertical shift factors have been added that make the distinction between clusters more obscure.

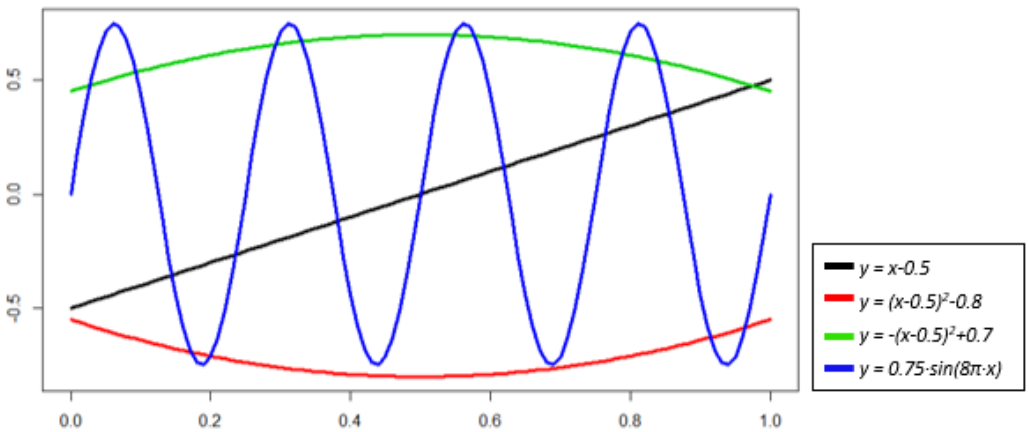


Fig. 3.33. Graphical representation of model 2.

Model 3

Model 3 tests fast variations on a periodic signal, but cannot be used to analyze the impact of peaks. The Gaussian kernel[7] will be used to simulate these sudden variations in a non-periodic function. This representation is convenient because the spread can be tuned in order to produce sharper or smoother peaks, following

$$f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the mean and σ^2 the spread. From the table we can see that this model has one more cluster than models 1 and 2. Additionally, clusters 4 and 5 have been mirrored horizontally and translated vertically so that the peak happens in the decreasing direction, as shown by the light blue and the dark blue lines depicted in Fig. 3.34.

Note that the *x*-axis now ranges from -10 to 10, and has a step of 0.1.

TABLE 3.12.
MODEL 3 MEAN AND SPREAD PARAMETERS.

Cluster Number	Mean	Spread	Function	Color
1	*0*	*2*	$y = \frac{1}{2\sqrt{2\pi}} \cdot e^{-\frac{(x)^2}{2 \cdot 2^2}}$	Black
2	*-2*	*1*	$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x+2)^2}{2 \cdot 1^2}}$	Red
3	*2*	*1*	$y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-2)^2}{2 \cdot 1^2}}$	Green
4	*0*	*1*	$y = \frac{-1}{\sqrt{2\pi}} \cdot e^{-\frac{(x)^2}{2 \cdot 1^2}} + 0.4$	Dark blue
5	*0*	*3*	$y = \frac{-2}{3\sqrt{2\pi}} \cdot e^{-\frac{(x)^2}{2 \cdot 3^2}} + 0.4$	Light blue

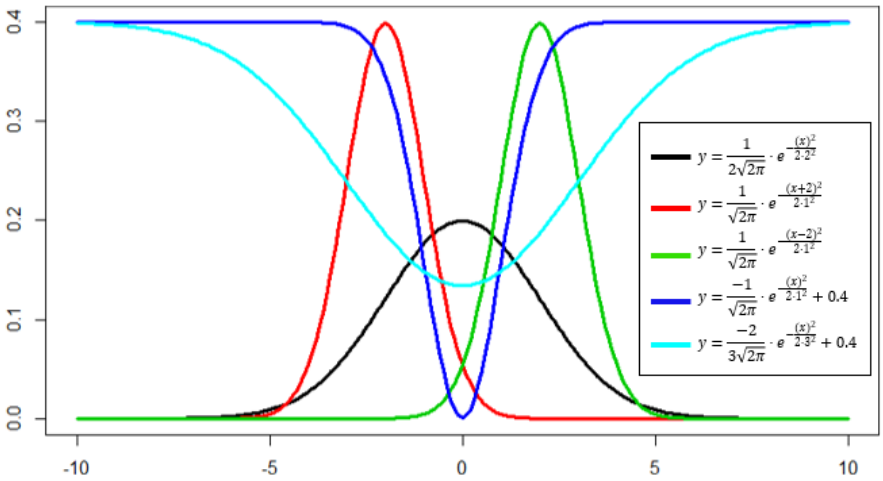


Fig. 3.34. Graphical representation of model 3.

[7] The Gaussian kernel is the unnormalized version of the Gaussian probability density function.

Model 4

This model is taken from Leroy et al., (2018) [44] and is used in the article to mimic swimmers' progression curves for clustering. The *x*-axis is defined in the interval [0,1] and with a step of 0.05. By taking this model from an external source we make sure that the clustering results obtained from the models are not biased.

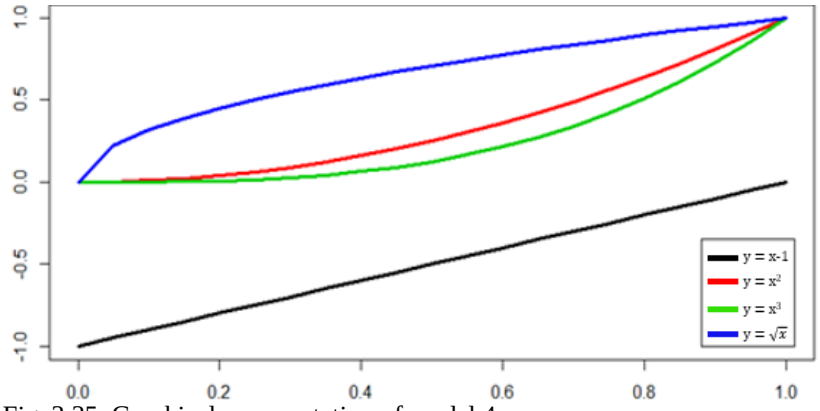


Fig. 3.35. Graphical representation of model 4.

3.4 Coefficient Clustering

A common alternative to clustering the values of the approximated function is to use the vector of coefficients that represent that function in order to find the final grouping. This is because two functions belonging to the same cluster are expected to have a small sum of squared distances to the centers determined by their coefficients.

In other words, two functions that are alike would be represented using a similar coefficient vector, and hence the K-Means clustering of coefficients is a convenient way of reaching a clustering result.

In most situations, the coefficient vector is considerably smaller in size than the functional data values vector. Thus, this clustering method is commonly more efficient in terms of execution time. In sections 4.1.2, 4.2.2, 4.3.2 and 4.4.2 we assess the quality of such clustering when using the coefficients as an input to KM, KMPP and MVMBD and compare it to that obtained by FMBD over the original input data (Table 4.27).

3.5 Missing Data

It is easy to analyze functions when all data are observed at the same time instants and when the observation interval is the same for all functions. However, when working with real datasets it is common to find that not all data has been collected at the same time instants, or that a value, at a certain time point, is missing for some function.

The term *missing data* is used to refer to the situation described in Fig. 3.36, where observations corresponding to red dots are missing. This can be addressed in several ways; the most common ones are:

- Vertical Analysis
  - *Mean of the column*. Substitute the missing values for the mean of that component over all data points
  - *Median of the column*. Same procedure as before but instead of using the mean, we choose the median.
- Horizontal Analysis
  - *Linear interpolation* between components to the left and to the right of the missing observation.

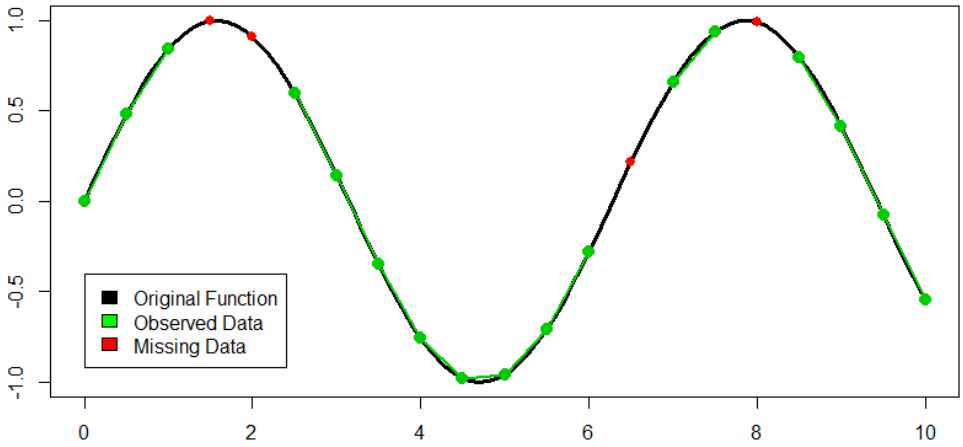


Fig. 3.36. Missing data.

Additionally, in our simulations, we assume that there are at least two observations for each function: its first and last values. For example, consider the data matrix in Table. 3.13.

TABLE 3.13.
MISSING VALUE DATA MATRIX.

Function Number	x_1	x_2	x_3	x_4	x_5	x_6
F1	1	5	NA	NA	17	21
F2	2	6	NA	14	18	22
F3	3	7	NA	15	19	23
F4	4	8	NA	16	20	24

NA stands for Not Available or, equivalently, that there is no data for that function in that coordinate.

After the interpolation, the resulting matrix is shown in Fig. 3.14.

TABLE 3.14.
INTERPOLATED VALUES USING THE THREE APPROACHES DESCRIBED

MEAN INTERPOLATION MEDIAN INTERPOLATION LINEAR INTERPOLATION

Function Number	x_1	x_2	x_3			x_4			x_5	x_6
F1	1	5	5	5	9	15	15	13	17	21
F2	2	6	6	6	10	14			18	22
F3	3	7	7	7	11	15			19	23
F4	4	8	8	8	12	16			20	24

In this case linear interpolation works the best as the obtained values are exactly the desired ones. In cases where there are large variations from one coordinate to the next, linear interpolation may not provide the best results and hence mean and median interpolation can be used instead.

In the context of this project, linear interpolation is used in order to simplify the results provided. This way, the project is focused on the evaluation of MBD-based initialization for functional data and not other supplementary concepts.

TABLE 3.15.
IMPLEMENTATION OF MISSING VALUES IN R.

Implementation in R
For testing purposes, the function `miss()` has been implemented. It receives `Pmiss` as an argument, representing the probability of having a missing observation in a function. The data matrix is created and the function `miss()` is run over it destroying (setting to NA) values with the given probability. As explained before, the first and the last observations of a function are always assumed to be present, so these values are always kept after running `miss()`.

3.6 Other Limitations

In this section, a signal processing perspective will be acquired to analyze some considerations about processing functional data. In particular, B-Splines will be carefully explored in order to understand the limitations of the curve fitting method proposed.

Different Sampling Rates of the Input Signal

The data to be clustered is obtained by sampling a continuous function at specific time instants. Therefore, different sampling rates of the same function produce different data sets.

Consider a noiseless sine wave, with unit angular frequency and amplitude. The function has been observed in the interval [0,100] at two sampling periods: $T_1 = 1$ and $T_2 = 2$.

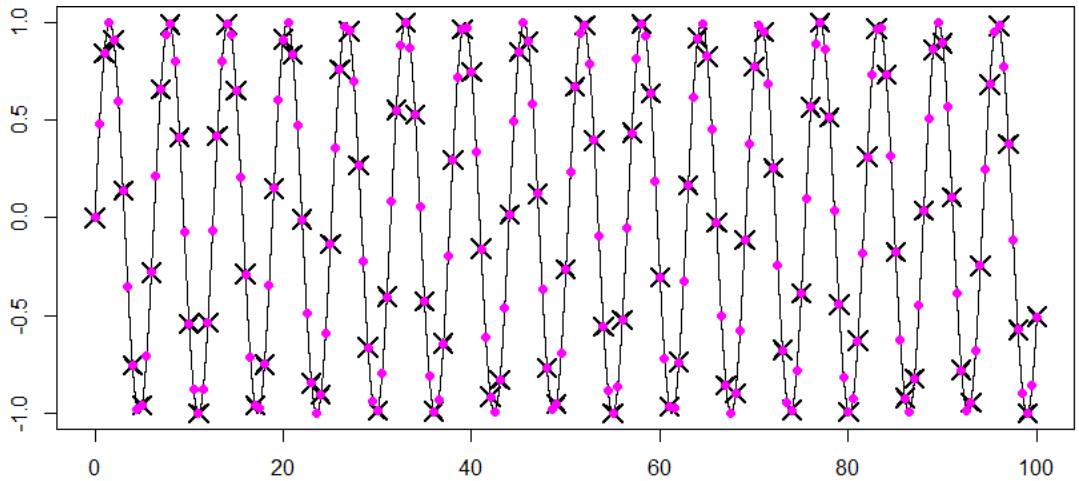


Fig. 3.37. Different sampling frequencies: magenta for $T_1 = 1$ and black for $T_2 = 2$.

Fig. 3.37 shows the original waveform (black lines) with the observations. The black dots correspond to the larger sampling period's samples ($T_2 = 2$). Having a smaller sampling period means that we have more observations (magenta crosses).

It is convenient to use B-Splines with relatively high degrees of freedom to account for all the oscillations that the sine introduces: the fast variations between -1 and 1 mean that a better fit is required. With the same knots - or equivalently, degrees of freedom - for both cases (T_1 and T_2) we have different approximation results.

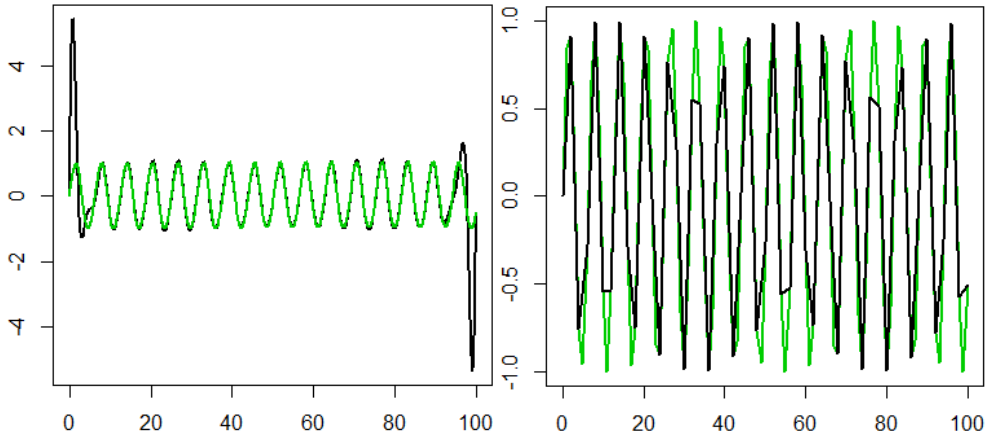


Fig. 3.38. Function approximation for different sampling rates.

The graph on the left of Fig. 3.38 shows the result of the B-Spline function approximation method. The green curve represents the result for the case $T_1 = 1$, while the black curve represents the B-Spline interpolation fit for $T_2 = 2$. The graph on the right shows the linear interpolation of samples for comparison, with the same color code.

Although the degrees of freedom for fitting the function are large, it can be seen that they are not high enough to approximate accurately the data set with less original samples ($T_2 = 2$). This can be seen by the presence of peaks at the interval extremes in the black curve of the left graph in Fig. 3.38.

On the other hand, there are two special considerations that must be kept in mind:

1. *The sampling frequency satisfies Nyquist's theorem* [19].
   The angular frequency of the sine is 1, which means that the sampling period must be at most equal to π. Both data sets, for $T_1 = 1$ and $T_2 = 2$, satisfy this condition.
2. *Other function approximation techniques can be used.*
   Fourier analysis can be considered in this case as the function exemplified is periodic.

Are sampling rates a problem?

Different sampling rates can be a problem even if Nyquist's theorem is satisfied. This issue becomes relevant when programming. Different sampling rates entail having datasets of different sizes, which makes it harder to store them into a matrix.

In the example above, the first dataset is twice as big as the second one because one sampling period is half the other. This can be accounted for in the approximation via the oversampling factor (OSF). Having an $OSF_2 = 2 \cdot OSF_1$ ensures that the number of output samples of the approximated function for the second dataset is twice the number of samples of the input, raw data, for that dataset. Mathematically,

$$A_i = N_i \cdot OSF_i \qquad N_1 = 2 \cdot N_2 \qquad OSF_1 = \frac{1}{2} \cdot OSF_2$$

$$\therefore A_1 = N_1 \cdot OSF_1 = 2 \cdot N_2 \cdot OSF_1 = 2 \cdot N_2 \cdot \frac{1}{2} \cdot OSF_2 = N_2 \cdot OSF_2 = A_2$$

$$\therefore A_1 = A_2 \ \blacksquare$$

where, N_i is the number of samples in dataset i, and A_i is the number of samples of the approximated function i.

But what if we have a large number of datasets with different observations at different time instants?

The `oversample()` function has been programmed in R to receive the data matrix and the B-Spline approximation arguments to perform the function approximation processing, and to output a data matrix with the new approximated values for each datapoint.

This problem of different dataset dimensions can be tackled by the use of the `vec` parameter in the `oversample()` function. This parameter allows passing the desired time points (*x*-axis values) in which the new, approximated datapoints, will be observed.

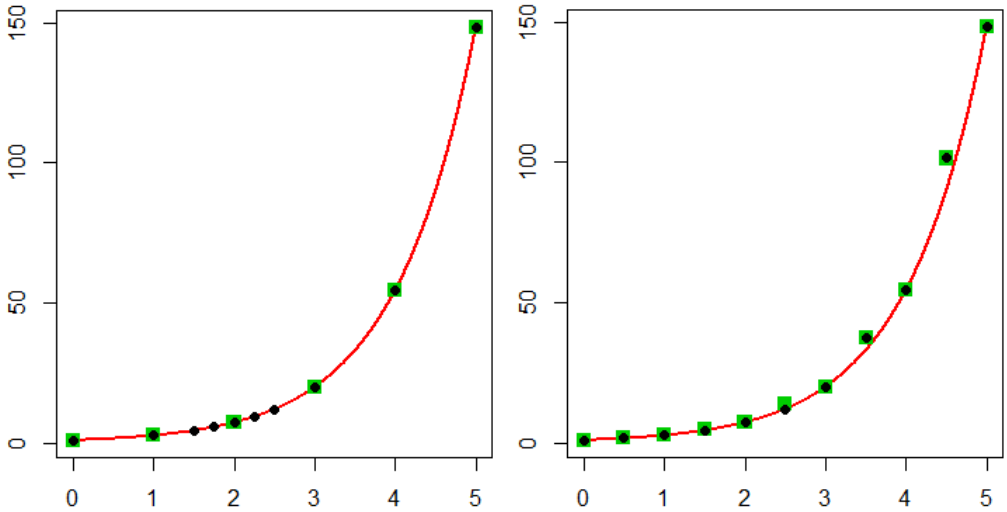


Fig. 3.39. Exponential function with different observation sets.

The graph on the left of Fig. 3.39 shows a first data set (represented as black dots), with more observations than a second data set (green squares). In fact, in this case, the second dataset is a subset of the first one.

The degrees of freedom and the degree of the polynomials are set to an appropriate level (*DF* = 50 in this case), taking into account the size of the datasets, to produce the graph on the right. This graph represents the outputs produced by the `oversample()` function for the two datasets.

The `vec` parameter has been set to have a step of 0.5 and a range of 0 to 5. The output dataset consists of other values than the original one, but has the same number of observations at the same time instants. The same color code is used in both panels.

The Observation Interval

After addressing the problem of having different dataset sizes due to different sampling rates the issue of different observation interval lengths has to be considered. Although we may have the same number of samples, the time interval in which these are obtained from may not be the same. How can we apply the oversampling function to address this matter?

Let us consider two datasets corresponding to a quadratic polynomial. The first dataset is observed in the interval [0, 4], which corresponds to the black dots in Fig. 3.40, while the second one is observed in the interval [0, 3.5], represented in green squares (left panel). Both of them have the same sampling period, 0.5.

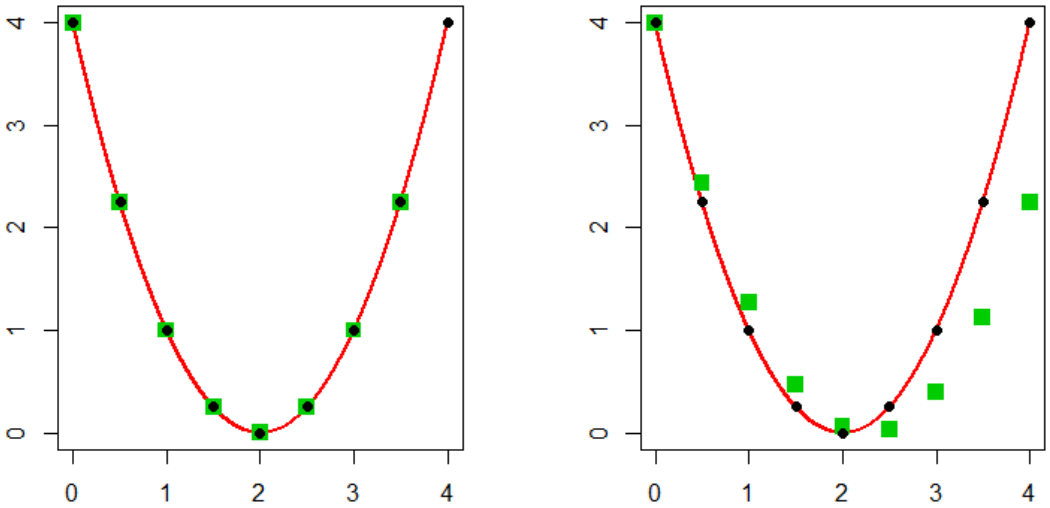


Fig. 3.40. Truncated observation interval and approximation for a quadratic function.

It can be seen that the first data set is larger than the second one. The original function that is sampled is represented in red. The graph on the left in Fig. 3.40 shows the input samples and the one on the right displays the B-Spline-approximated output samples (in black and green respectively). The green dataset, that has been observed in a smaller time interval, results in a worse approximation (right panel).

By changing the sampling frequency to a higher value and keeping the same observation interval we can see that the approximation results do not change. This means that all the relevant information is extracted with a smaller sampling frequency.

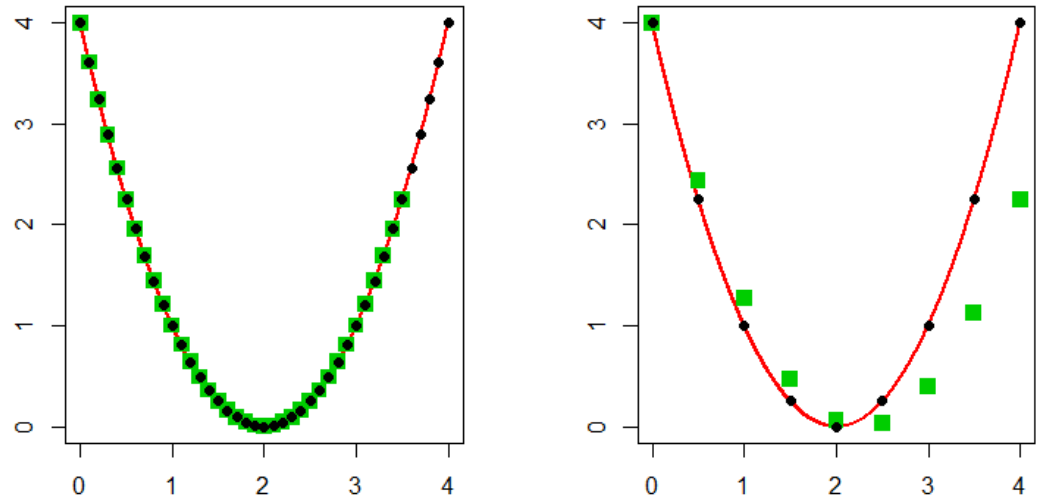


Fig. 3.41. Higher sampling rate on a truncated observation interval.

However, we can consider a different definition following the same example. Let us make both datasets the same, except for the last sample, which will not be present in the second set.

In Fig. 3.42. we can see that the B-Spline approximation output samples (green dots) are now adjusted more to the input curve. Hence, the more samples we have inside the interval in which the approximation of the function is performed, the better the approximation will be.

This has been tested with other curves, not just a parabola, and the same finding has been observed.

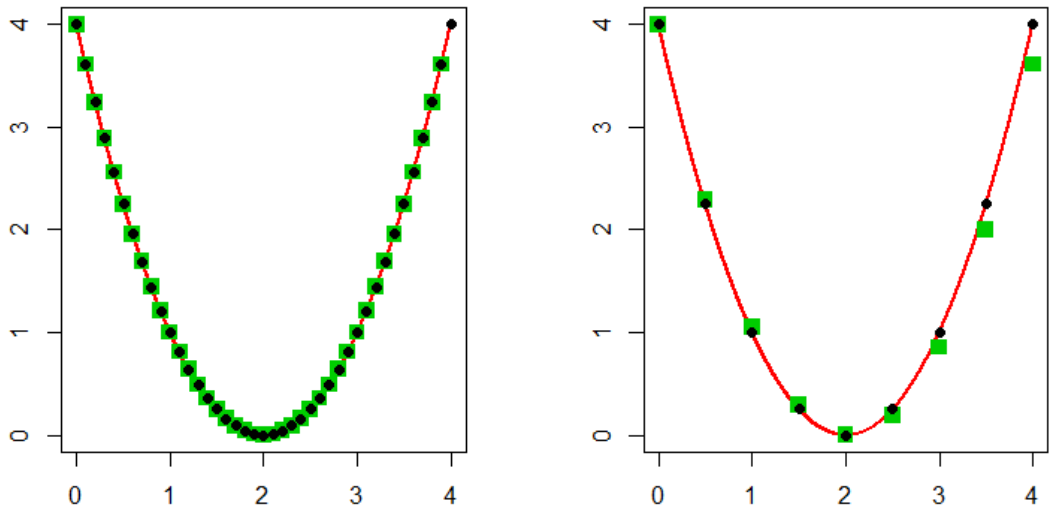


Fig. 3.42. Missing last sample on a B-Spline approximation of a quadratic function.

Samples Outside the Observation Interval

We have seen that a longer observation interval provides a tighter fit of the output samples to the original curve.

It can be stated that the B-Spline function approximation technique is not able to *predict* accurately the shape of the curve from the given samples at positions outside the observation interval. It is useful, however, in the generation of samples inside the observed range. For prediction purposes other estimation techniques, which are currently not in place, can be used (mathematical models [20], Kalman filtering [45], etc.).

Additionally, other techniques of improving curve fitting will be discussed in Section 7 (Future Studies).

3.7 Real Data

Testing MBD-based initialization of K-Means on the models proposed does not translate into a direct application to real world scenarios. Real data, carefully chosen to meet the privacy requirements, is used to evaluate the results in these cases and assess whether the proposed algorithm is of practical use.

Continuing the reasoning followed in Section 3.3 (*Models for Testing*), model 1 was conceived to loosely represent temperature curves in different climates.

Worldwide weather data is obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program [46]. It includes meteorological data, amongst others, on temperature and precipitation.

Specially, different geographical zones of three different climates will be considered, as shown in Fig. 3.43. Each data point consists of 365 coordinates associated to the daily measurements of 2018. Table 3.16 summarizes the data analyzed.

Climate classifications have been done according to the standard known as the Köppen system [47]. Table 3.17 states the equivalence between the terms used in the project and the group code of the classification system.

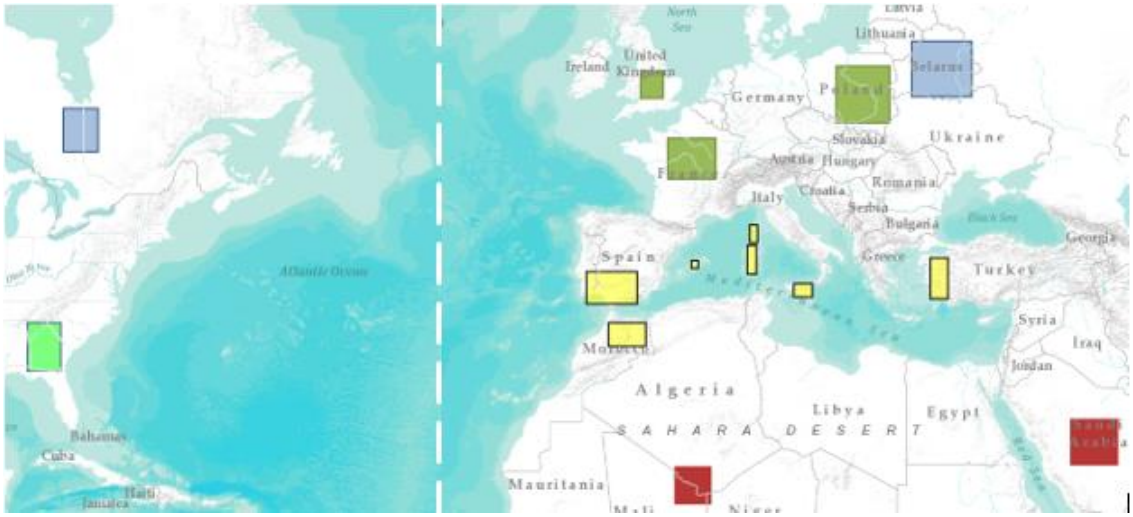


Fig. 3.43. Approximate regions for climate data collection.

TABLE 3.16.
METEOROLOGICAL DATA BY CLIMATE.

Region	Latitude		Longitude		Number of points	Climate
	From	To	From	To		
Belarus	52.75	54.75	26.75	30.75	45	Savanna
Corsica	41.75	42.75	8.75	9.25	6	Mediterranean[8]
England	51.25	52.75	-2.25	-0.25	20	Oceanic
France	46.75	49.25	-0.25	4.25	60	Oceanic
Georgia, US	30.25	34.75	-85.75	-81.25	100	Humid Subtropical
James Bay, Canada	48.25	50.75	-81.25	-76.25	66	Savanna
Majorca	39.25	39.75	2.75	3.25	4	Mediterranean[8]
North Morocco	34.25	35.25	-6.25	-2.25	30	Mediterranean[8]
Poland	51.25	53.75	19.75	24.25	60	Oceanic
Sahara	18.25	21.25	1.25	3.75	42	Hot desert
Sardinia	39.25	40.75	8.75	9.75	12	Mediterranean[8]
Saudi Arabia	22.25	26.75	41.75	46.25	100	Hot desert
Sicily	37.25	38.25	12.75	15.25	18	Mediterranean[8]
Southern Iberian Peninsula	37.25	38.75	-8.25	-3.25	44	Mediterranean[8]
Western Turkey	36.75	39.75	27.25	28.75	28	Mediterranean[8]

[8] Mediterranean climate refers to hot summer Mediterranean climate.

TABLE 3.17.
KÖPPEN CLIMATE CLASSIFICATION EQUIVALENCE.

Climate	Köppen climate classification system group code
Savanna	Aw
Hot Dessert	BWh
Oceanic	Cfb
(Hot Summer) Mediterranean	Csa
Humid Subtropical	Cwa

3.8 Proposed Method as an R Package

Community content is an essential aspect of R, and part of the coding done in this project has been eased thanks to shared content. Therefore, having developed the code for MBD-based initialization of K-Means for functional data in several R scripts, it is easy to assemble it into an R package. A suitable documentation has to come along with the code to make it accessible for research purposes.

Conversely, as the programming of an application is not the main goal of the project, no further inspection will be done in technicalities of the documentation. Section 7 (*Future Studies*) discusses this matter in more detail.

4. RESULTS

In this section, a descriptive perspective of the findings will be considered. The clustering outputs are appraised according to the evaluation techniques discussed in the *Methods* chapter (subsection 3.1.4) and results are presented in tables and figures accordingly.

To that end, we use the four models presented with different characteristics, as well as real data. All these testing situations provide an exhaustive approach to the analysis of the proposed method.

Moreover, a telecommunications-based approach is considered, especially relevant at physical layer in the signal processing field. This insight is provided in Section 5 followed by a deeper discussion on filtering.

Our Hypothesis

As a continuation of the fundamental lines of work established in the Introduction (section 1.4), the subsequent question arises:

Is MBD-based initialization of K-Means for data approximated by B-Splines (FMBD) reliable for clustering?

Our hypothesis is that FMBD is an advantageous solution for centroid-based clustering, offering robust and consistent results in time-series datasets.

To test so, the parameters relevant to the B-Splines function approximation and bootstrapping will be tuned and tested in various noise scenarios. These are collected in Table 4.1.

TABLE 4.1.
PARAMETERS CONSIDERED IN THE EXPERIMENTS.

Parameter	Name	Function
Intercept	`Intercept`	B-Splines
Degree of polynomial	`DP`	
Degrees of freedom	`DF`	
Oversampling Factor	`OSF`	
Bootstrapping replicas	`B`	MBD
Number of simulations	`N`	Simulation parameters
Number of clusters	`Nclus`	
Noise standard deviation	`Sigma`	Models

Testing the Hypothesis

In order to determine whether our prediction on FMBD being an advantageous solution for initializing K-Means is true or not, models 1 through 4 are used as well as real data. For each model, the Correctness, Purity, ARI, Distortion, Iterations and Execution Time (CoPADIT) measures are calculated for K-Means (KM), K-Means with multivariate MBD initialization (MVMBD), K-Means with MBD initialization for functional data (FMBD), K-Means Plus Plus (KMPP) and K-Means Plus Plus over functional data (FKMPP). Additionally, coefficient clustering and missing data cases are tested, and the statistics and distributions of the CoPADIT measures are provided. The execution time is measured with the hardware and software specifications detailed in Table 4.2.

TABLE 4.2.
HARDWARE AND SOFTWARE SPECIFICATIONS

	Component	Specifications
Software	Operating System	*Windows 10 Pro (64bits)*
	IDE	*RStudio*
Hardware	CPU	*Intel Core i7-6700HQ*
	RAM	*8 GB*
	GPU	*NVIDIA GeForce 960M*

Furthermore, the parameters used to perform the function approximation using B-Splines that appear in the various tables of this section (Tables 4.3, 4.9, 4.14, 4.19 and 4.24) have been chosen after intensive research, as presented in the *OSF and DF Behavior According to Input Data* section in the Appendix. The choice of these parameters yields the best B-Spline approximation to the input data for clustering purposes.

For the sake of clarity, the results provided in this section were collected using an archetypal value of the noise standard deviation, `sigma` = 1. Please refer to the Appendix at the end of the report to find the results for values of `sigma` = 0, 0.5, 1, 1.5 and 2, along with the corresponding p-values of the t-test for the equality of means. The coefficient clustering results for `sigma` = 1 and 2 are provided as well.

4.1 Model One

Table 4.3. shows the parameters that provide the best results for the function approximation for clustering purposes in different noise scenarios.

TABLE 4.3.
MODEL 1 OPTIMAL PARAMETERS.

Parameter	Value
Intercept	*True*
Degree of polynomial	*3*
Degrees of freedom	*4*
Oversampling Factor	*1*

The five-way comparison, the coefficient clustering and missing data experiments are carried out with the values from Table 4.3. There are 4 clusters in this model, and the number of functions per cluster is set to 25. The bootstrapping factor, *B*, is set to 25 as a standard value for all models and simulations are repeated 1000 times to get results with statistical significance. Results are reported in the following tables and figures.

4.1.1 Five-Way Comparison

The mean, median and variance provide a summary for the statistics. In this case, `sigma` = 1.

TABLE 4.4.
MODEL 1 SUMMARY STATISTICS, 5-WAY COPADIT FOR SIGMA = 1.
MEDIAN, MEAN AND VARIANCE FOR EACH METHOD

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.66	0.67	0.4294	9568	4	~ 0
	Mean	0.6524	0.667	0.42	9572	3.721	1.032e-03
	Variance	0.005021	0.003507	0.007326	20130	0.5938	6.777e-06
MV MBD	Median	0.67	0.68	0.4369	9575	3	0.08278
	Mean	0.6596	0.6718	0.4315	9573	3.484	0.08855
	Variance	0.004463	0.003272	0.007274	20290	0.6684	0.0004869
FMBD	Median	0.84	0.84	0.6513	9650	2	0.1862
	Mean	0.8253	0.8277	0.6467	9651	1.949	0.1955
	Variance	0.003671	0.002979	0.005546	20170	0.1125	0.001486
KMPP	Median	0.65	0.67	0.4265	9684	4	2.992e-03
	Mean	0.6487	0.6675	0.4193	9687	3.760	3.993e-03
	Variance	0.00466	0.003513	0.007457	20780	0.563	1.947e-05
FKMPP	Median	0.74	0.8	0.6103	9656	3	0.1089
	Mean	0.7404	0.795	0.6099	9659	2.764	0.114
	Variance	0.01134	0.004935	0.008153	20390	0.4387	0.0006833

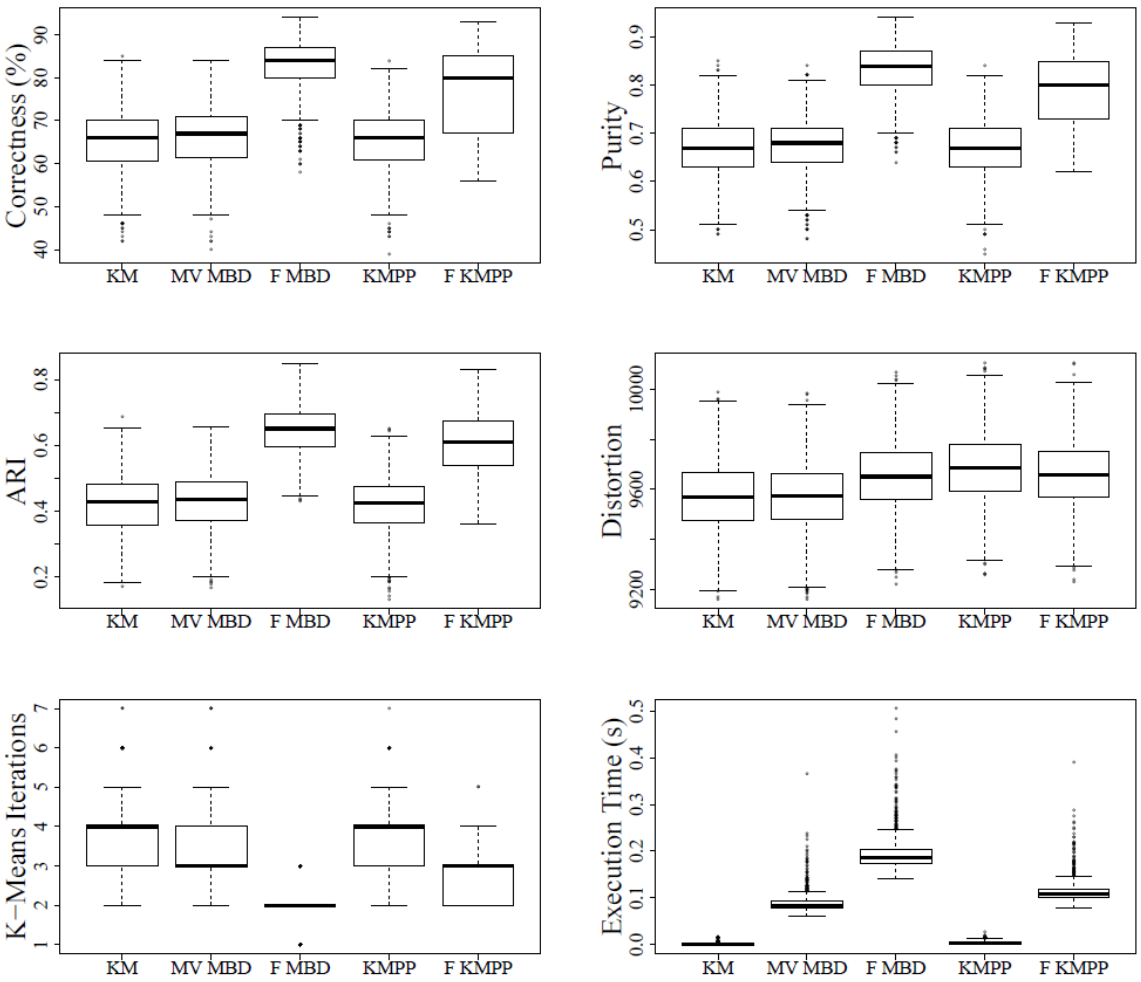


Fig. 4.1. Model 1, 5-way distribution of CoPADIT measures for sigma = 1.

The p-values for the paired t-test for equality of means of correctness, purity, and ARI for all methods are collected in Table 4.5.

TABLE 4.5.
MODEL 1 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

sigma = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	5.470e-03	~ 0	9.105e-01	2.290e-160
	Purity		1.763e-02	~ 0	8.227e-01	6.015e-256
	ARI		3.219e-05	~ 0	8.196e-01	1.333e-293
MV MBD	Correctness	5.470e-03	-	~ 0	7.693e-03	4.447e-144
	Purity	1.763e-02		~ 0	3.246e-02	1.084e-242
	ARI	3.219e-05		~ 0	1.211e-05	7.620e-274
FMBD	Correctness	~ 0	~ 0	-	~ 0	4.483e-50
	Purity	~ 0	~ 0		~ 0	1.639e-41
	ARI	~ 0	~ 0		~ 0	8.155e-40
KMPP	Correctness	9.105e-01	7.693e-03	~ 0	-	1.793e-158
	Purity	8.227e-01	3.246e-02	~ 0		6.214e-252
	ARI	8.196e-01	1.211e-05	~ 0		9.524e-294
FKMPP	Correctness	2.290e-160	4.447e-144	4.483e-50	1.793e-158	-
	Purity	6.015e-256	1.084e-242	1.639e-41	6.214e-252	
	ARI	1.333e-293	7.620e-274	8.155e-40	9.524e-294	

The test for equality of medians as implemented in the `Median.test()` function in R produces the results collected in Table 4.6. This test is performed to provide a more robust proof than the t-test for the equality of both distributions, accounting for the atypical data present in the CoPADIT measures.

TABLE 4.6.
MODEL 1 FMBD ACCURACY MEASURES' P-VALUE FOR THE EQUALITY OF MEDIANS TEST FOR SIGMA = 1.

`sigma` = 1							
Method			KM	MVMBD	FMBD	KMPP	FKMPP
FMBD	Correctness	p-value	2.086e-290	4.415e-286	-	7.331e-292	4.590e-12
	Purity	p-value	4.243e-291	9.018e-287		1.489e-292	4.590e-12
	ARI	p-value	1.482e-323	3.790e-290		1.246e-320	1.139-13

Similar results are obtained for the rest of the models for the pairwise comparisons of medians. The tables are not reported for simplicity and due to their similarity to the p-values of the t-test.

4.1.2 Coefficient Clustering

The results we obtain after clustering the B-splines coefficients with KM, MVMBD and KMPP for `sigma` = 1 are summarized in Table 4.7 and in the boxplots shown in Fig. 4.2.

TABLE 4.7.
MODEL 1 SUMMARY STATISTICS FOR COEFFICIENT CLUSTERING, 3-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.55	0.57	0.2589	9821	3	0.1017
	Mean	0.548	0.5724	0.2662	9828	2.715	0.1109
	Variance	0.006042	0.00366	0.006465	23860	0.4302	0.001383
MV MBD	Median	0.55	0.57	0.2582	9824	2	0.1131
	Mean	0.5451	0.5695	0.2637	9828	1.949	0.1227
	Variance	0.005907	0.003722	0.006438	23970	0.1045	0.001688
KMPP	Median	0.54	0.56	0.2415	9833	2	0.1043
	Mean	0.5358	0.5637	0.255	9834	2.492	0.1131
	Variance	0.006122	0.003627	0.006269	23550	0.3863	0.001452

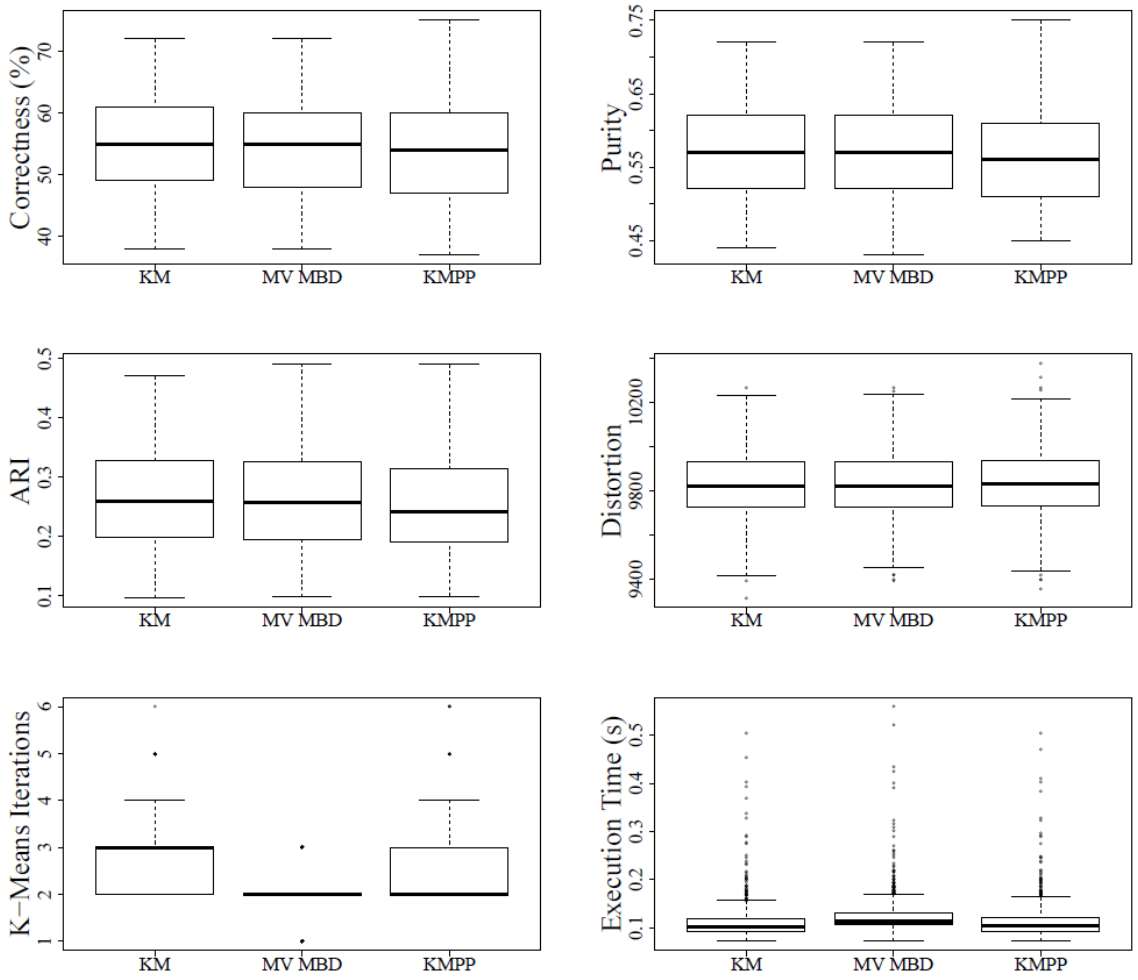


Fig. 4.2. Model 1, 3-way distribution of CoPADIT measures for sigma = 1.

We see that the three methods produce similar results. When compared to FMBD, none of the coefficient clustering techniques is capable of yielding better results for correctness, purity, ARI or distortion.

4.1.3 Missing Data

The results obtained by performing clustering on models with 25% missing data for `sigma` = 1 are summarized in Table 4.8 and in the boxplots shown in Fig. 4.3. We can see that the most accurate method is FMBD. Similar conclusions can be reached for percentages of missing data of 50% and 75% (please refer to the Appendix).

TABLE 4.8.
MODEL 1 SUMMARY STATISTICS FOR 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

`Pmiss` = 0.25, `sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.64	0.65	0.3825	8427	4	0.217
	Mean	0.638	0.6516	0.381	8429	3.653	0.2248
	Variance	0.004535	0.003137	0.0062	22480	0.5131	0.0009733
MV MBD	Median	0.65	0.66	0.3973	8430	3	0.2974
	Mean	0.6434	0.6564	0.3948	8432	3.336	0.3088
	Variance	0.003532	0.002422	0.005411	22460	0.6878	0.001452
FMBD	Median	0.78	0.78	0.5537	8513	2	0.1896
	Mean	0.769	0.7732	0.5516	8508	2.001	0.1967
	Variance	0.004472	0.003392	0.005669	22880	0.1051	0.0009345
KMPP	Median	0.64	0.65	0.3817	8426	4	0.22
	Mean	0.6333	0.648	0.3782	8429	3.651	0.2288
	Variance	0.004288	0.002807	0.005614	22220	0.5477	0.001066
FKMPP	Median	0.74	0.74	0.5182	8518	3	0.1108
	Mean	0.726	0.7457	0.5219	8516	2.776	0.1164
	Variance	0.007462	0.003933	0.006164	23250	0.3782	0.0004766

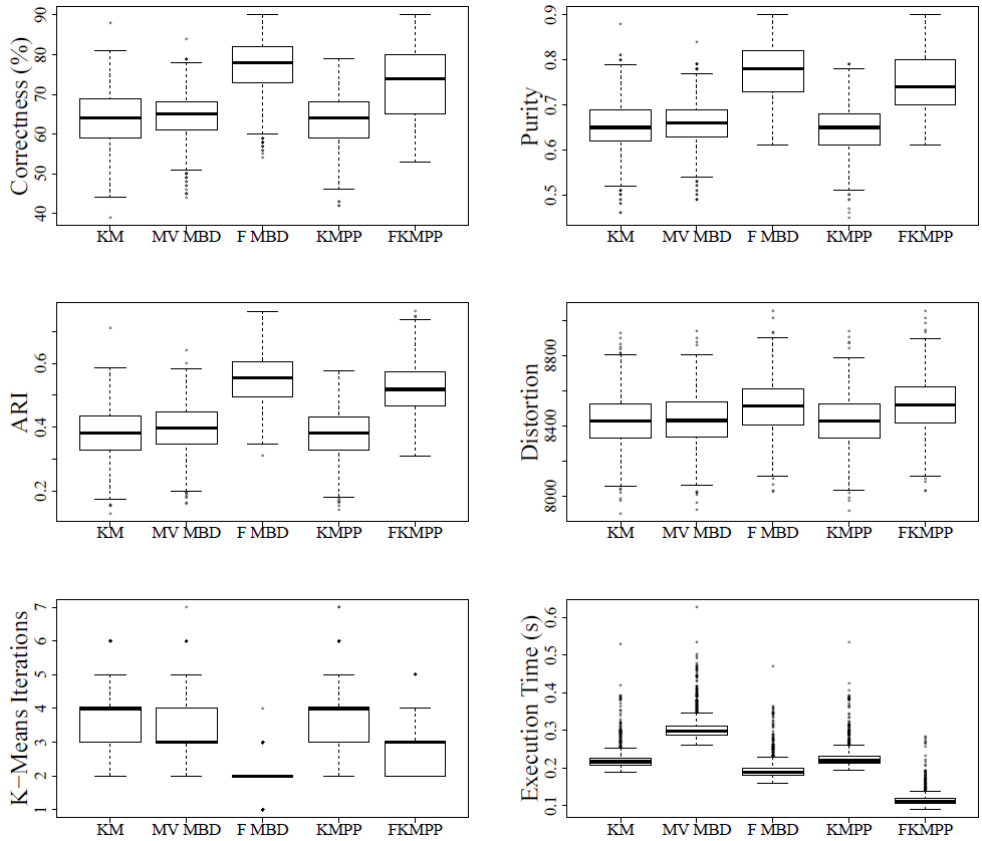


Fig. 4.3. Model 1, 5-way distribution of CoPADIT measures for sigma = 1 with 25% missing values.

4.2 Model Two

Table 4.9 shows the parameters that provide the best results for the function approximation for clustering purposes in different noise scenarios. The five-way comparison, the coefficient clustering and missing data results are presented using the values from Table 4.9.

TABLE 4.9.
MODEL 2 OPTIMAL PARAMETERS.

Parameter	Value
Intercept	*True*
Degree of polynomial	*3*
Degrees of freedom	*15*
Oversampling Factor	*1*

4.2.1 Five-Way Comparison

The results for `sigma` = 1 are summarized in Table 4.10 and in Fig. 4.4.

TABLE 4.10.
MODEL 2 SUMMARY STATISTICS, 5-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	1	1	1	9744	2	~ 0
	Mean	0.9236	0.9444	0.9137	9799	2.518	7.927e-04
	Variance	0.02038	0.01062	0.02481	70330	0.31	3.412e-06
MV MBD	Median	1	1	1	9683	1	0.07834
	Mean	0.9983	0.9984	0.9961	9687	1.301	0.08138
	Variance	0.0001429	7.61e-05	0.000243	21060	0.2166	0.0002188
FMBD	Median	1	1	1	9684	1	0.2251
	Mean	0.9994	0.994	0.9984	9687	1.023	0.2378
	Variance	5.934e-06	5.93e-06	4.3e-05	20780	0.02249	0.001664
KMPP	Median	1	1	1	9755	2	3.036e-03
	Mean	0.91	0.9341	0.8986	9821	2.455	3.783e-03
	Variance	0.02321	0.01228	0.02825	76280	0.2943	1.219e-05
FKMPP	Median	1	1	1	9735	2	0.1502
	Mean	0.934	0.9522	0.927	9791	2.09	0.1611
	Variance	0.01847	0.009576	0.02195	68310	0.2722	0.001262

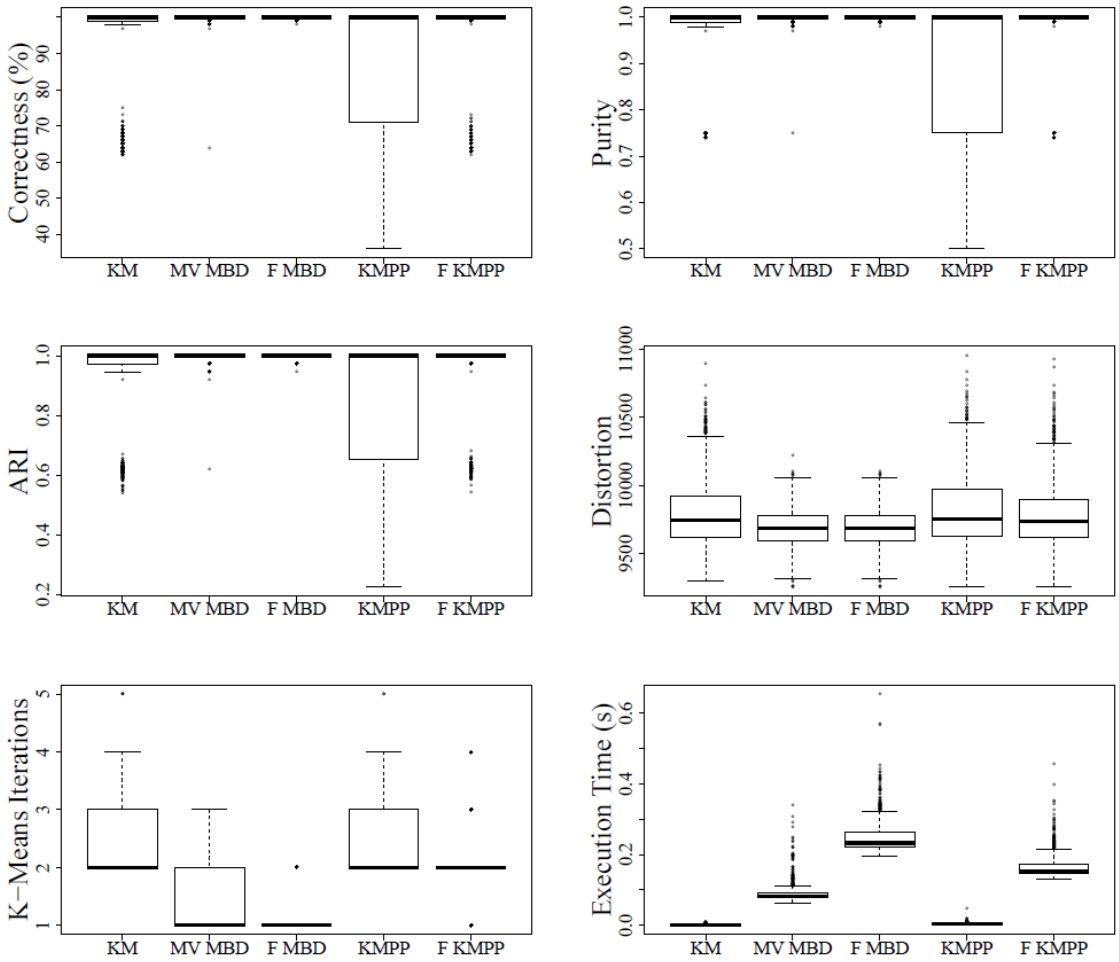


Fig. 4.4. Model 2, 5-way distribution of CoPADIT measures for sigma = 1

Although the medians for correctness, purity and ARI are the same and their equality is not rejected by the median test, the asymmetry in the distribution is accounted for in the t-test. The corresponding p-values for all methods are collected in Table 4.11.

TABLE 4.11.
MODEL 2 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

sigma = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	3.151e-54	5.144e-56	3.869e-02	1.033e-01
	Purity		1.566e-54	1.302e-56	3.159e-02	8.992e-02
	ARI		2.046e-54	2.282e-57	3.773e-02	5.752e-02
MV MBD	Correctness	3.151e-54	-	4.426e-03	5.017e-65	2.134e-45
	Purity	1.566e-54		4.265e-04	4.087e-65	2.869e-45
	ARI	2.046e-54		1.707e-06	5.775e-65	4.246e-44
FMBD	Correctness	5.144e-56	4.426e-03	-	2.472e-66	3.481e-47
	Purity	1.302e-56	4.265e-04		9.908e-67	2.104e-47
	ARI	2.282e-57	1.707e-06		1.459e-67	2.577e-47
KMPP	Correctness	3.869e-02	5.017e-65	2.472e-66	-	1.374e-04
	Purity	3.159e-02	4.087e-65	9.908e-67		7.364e-05
	ARI	3.773e-02	5.775e-65	1.459e-67		4.232e-05
FKMPP	Correctness	1.033e-01	2.134e-45	3.481e-47	1.374e-04	-
	Purity	8.992e-02	2.869e-45	2.104e-47	7.364e-05	
	ARI	5.752e-02	4.246e-44	2.577e-47	4.232e-05	

4.2.2 Coefficient Clustering

The results we obtain after clustering the B-splines coefficients with KM, MVMBD and KMPP for `sigma` = 1 are summarized in Table 4.12 and in the boxplots shown in Fig. 4.5. Again, a comparison of FMBD with the other methods demonstrates that it has a better performance in terms of the accuracy measures and distortion, while surpassed in execution time.

TABLE 4.12.
MODEL 2 SUMMARY STATISTICS FOR COEFFICIENT CLUSTERING, 3-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.77	0.77	0.6268	10290	3	0.1312
	Mean	0.7741	0.7859	0.6299	10300	3.155	0.1411
	Variance	0.006251	0.004161	0.008532	63700	0.4574	0.001368
MV MBD	Median	0.78	0.78	0.6352	10260	2	0.1521
	Mean	0.7912	0.7959	0.6432	10260	2.139	0.16
	Variance	0.004404	0.003566	0.006472	50580	0.1779	0.00157
KMPP	Median	0.77	0.77	0.6272	10290	3	0.1368
	Mean	0.775	0.7868	0.6316	10290	3.116	0.1429
	Variance	0.00569	0.003654	0.007323	57720	0.477	0.001401

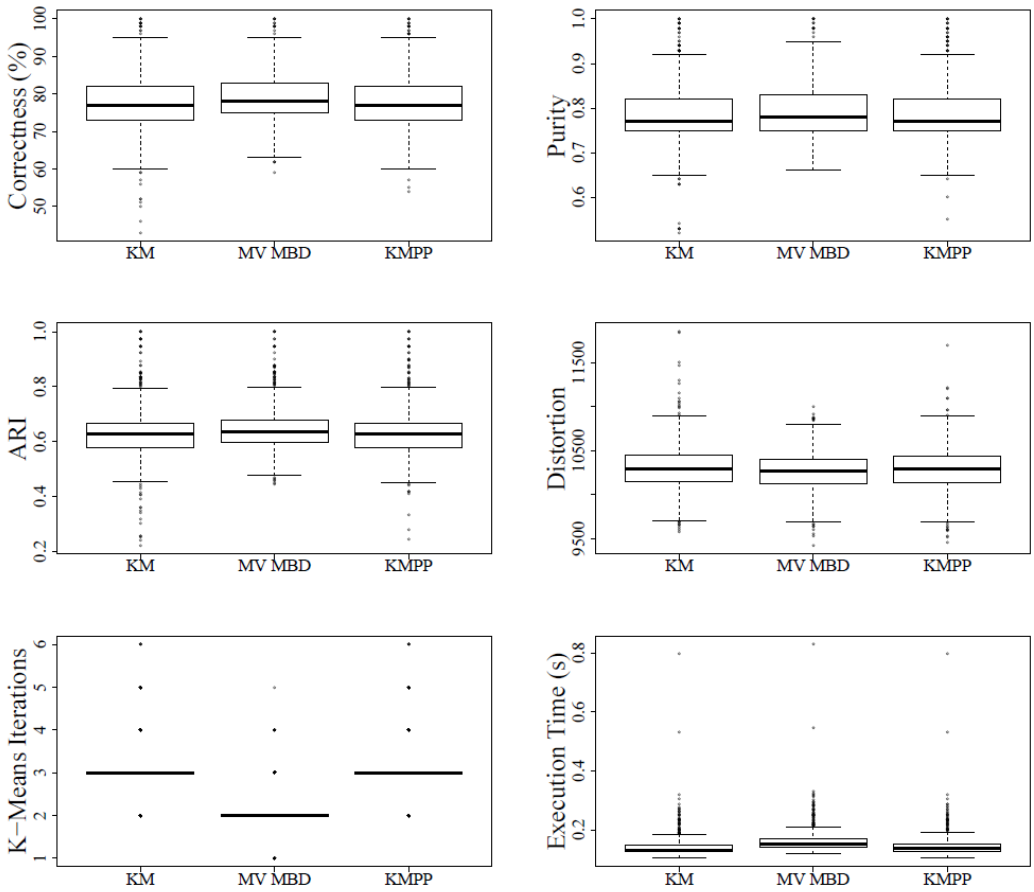


Fig. 4.5. Model 2, 3-way distribution of CoPADIT measures for `sigma` = 1

4.2.3 Missing Data

The results obtained by performing clustering on models with 25% of missing data for `sigma` = 1 are summarized in Table 4.13 and in the boxplots shown in Fig. 4.6. FMBD has recurrently the best behavior in terms of accuracy, as opposed to distortion and execution time.

TABLE 4.13.
MODEL 2 SUMMARY STATISTICS FOR 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

`Pmiss` = 0.25, `sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.99	0.99	0.9731	8607	3	0.1952
	Mean	0.933	0.9493	0.9152	8660	2.635	0.2094
	Variance	0.01682	0.008974	0.0213	66530	0.3701	0.004659
MV MBD	Median	1	1	1	8572	1	0.265
	Mean	0.9938	0.9938	0.9835	8571	1.469	0.2835
	Variance	6,23E-02	6,23E-02	0.000442	23140	0.2593	0.008379
FMBD	Median	1	1	1	8574	1	0.2036
	Mean	0.9961	0.9961	0.9896	8572	1.131	0.2158
	Variance	4,04E-02	4,04E-02	0.000288	23150	0.114	0.005168
KMPP	Median	0.99	0.99	0.9731	8608	3	0.197
	Mean	0.9383	0.9531	0.9216	8650	2.564	0.2119
	Variance	0.01573	0.008439	0.01983	59670	0.3102	0.00479
FKMPP	Median	1	1	1	8606	2	0.134
	Mean	0.9419	0.9568	0.9298	8655	2.319	0.1437
	Variance	0.01584	0.008329	0.01949	63940	0.2695	0.002646

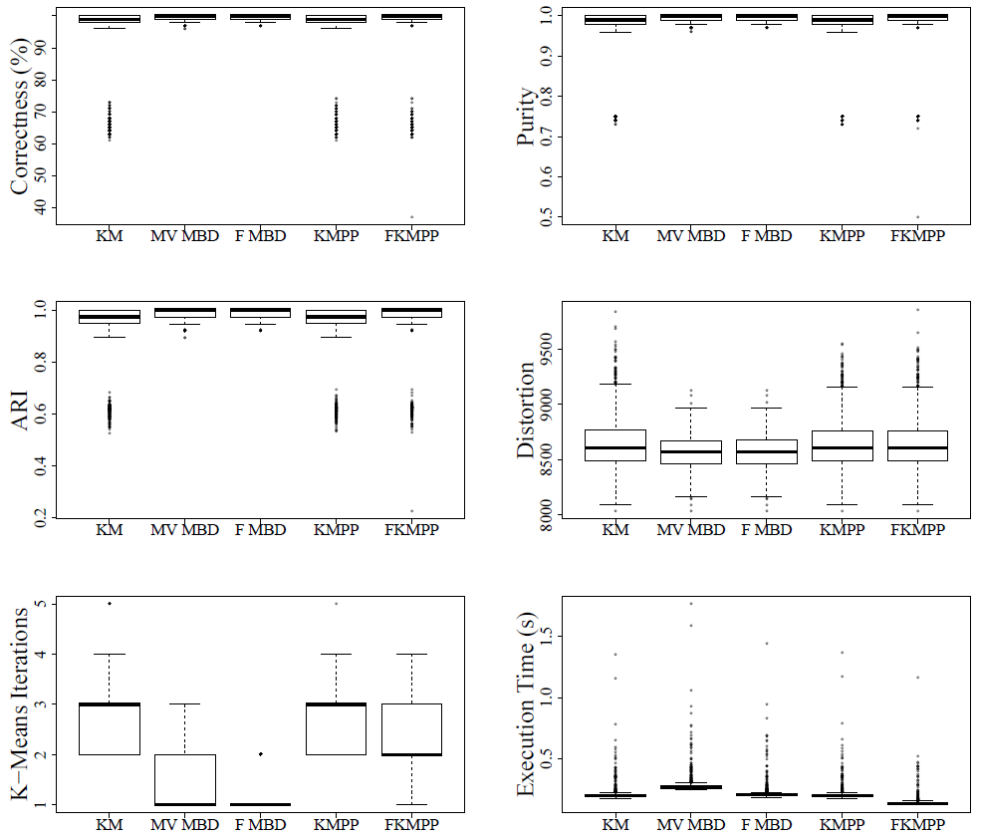


Fig. 4.6. Model 2, 5-way distribution of CoPADIT measures for `sigma` = 1 with 25% missing values.

4.3 Model Three

Table 4.14. shows the parameters that provide the best results for the function approximation for clustering purposes in different noise scenarios.

TABLE 4.14.
MODEL 3 OPTIMAL PARAMETERS.

Parameter	Value
Intercept	*True*
Degree of polynomial	*3*
Degrees of freedom	*13*
Oversampling Factor	*1*

The five-way comparison, the coefficient clustering and missing data results are presented using the values from Table 4.14.

4.3.1 Five-Way Comparison

The results for `sigma` = 1 are summarized in Table 4.15 and in Fig. 4.7. FMBD is also the best alternative for this model.

TABLE 4.15.
MODEL 3 SUMMARY STATISTICS, 5-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.44	0.448	0.2151	23680	4	1.994e-03
	Mean	0.4408	0.455	0.2176	23680	4.116	2.388e-03
	Variance	0.001321	0.001027	0.001423	52170	0.7273	8.160e-06
MV MBD	Median	0.432	0.452	0.2424	23680	4	0.2207
	Mean	0.4361	0.4535	0.243	23690	4.169	0.2312
	Variance	0.001049	0.000899	0.001571	50760	0.7672	0.001464
FMBD	Median	0.536	0.552	0.3336	23970	3	0.4112
	Mean	0.5382	0.5591	0.3336	23970	2.59	0.4244
	Variance	0.003123	0.002053	0.002033	52540	0.428	0.002837
KMPP	Median	0.44	0.448	0.2123	23680	4	0.01022
	Mean	0.4396	0.4533	0.215	23680	4.113	0.01238
	Variance	0.001263	0.000979	0.001298	51590	0.739	0.0001091
FKMPP	Median	0.52	0.552	0.3166	23980	3	0.2137
	Mean	0.5236	0.552	0.3175	23980	3.393	0.224
	Variance	0.003387	0.002134	0.002406	53700	0.465	0.001465

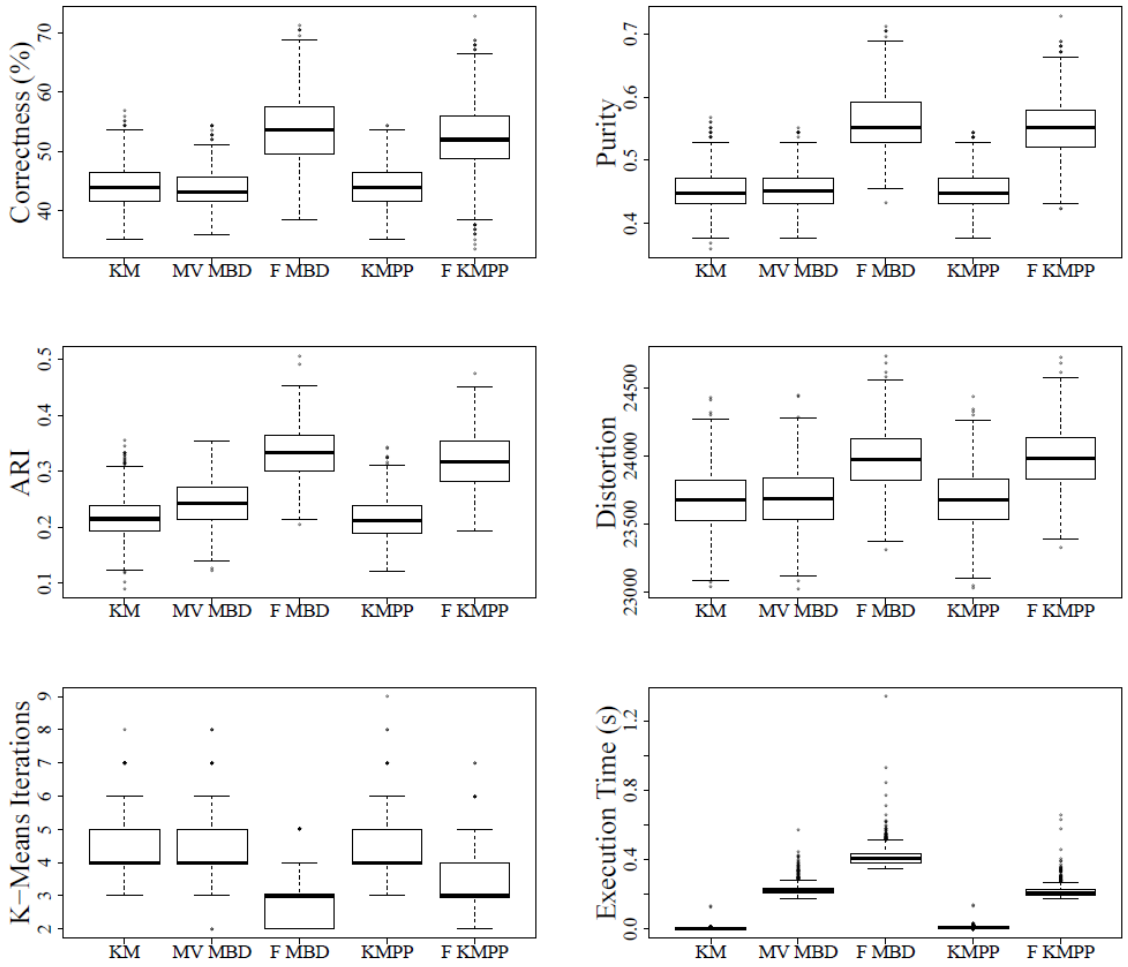


Fig. 4.7. Model 3, 5-way distribution of CoPADIT measures for `sigma` = 1

The p-values for the paired t-test for equality of means of correctness, purity, and ARI for all methods are collected in Table 4.16, with FMBD being significantly better than the other algorithms in terms of accuracy.

TABLE 4.16.
MODEL 3 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

`sigma` = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	1.375e-03	1.001e-262	4.345e-01	3.527e-204
	Purity		2.327e-01	~ 0	1.866e-01	9.288e-320
	ARI		3.874e-60	~ 0	5.608e-02	4.272e-297
MV MBD	Correctness	1.375e-03	-	9.213e-284	1.347e-02	1.082e-222
	Purity	2.327e-01		~ 0	0.8801	~ 0
	ARI	3.874e-60		5.042e-285	2.132e-72	1.426e-205
FMBD	Correctness	1.001e-262	9.213e-284	-	1.209e-258	2.920e-11
	Purity	~ 0	~ 0		~ 0	6.434e-06
	ARI	~ 0	5.042e-285		~ 0	3.481e-25
KMPP	Correctness	4.345e-01	1.347e-02	1.209e-258	-	7.438e-208
	Purity	1.866e-01	0.8801	~ 0		1.1818e-320
	ARI	5.608e-02	2.132e-72	~ 0		6.420e-306
FKMPP	Correctness	3.527e-204	1.082e-222	2.920e-11	7.438e-208	-
	Purity	9.288e-320	~ 0	6.434e-06	1.1818e-320	
	ARI	4.272e-297	1.426e-205	3.481e-25	6.420e-306	

4.3.2 Coefficient Clustering

The results we obtain after clustering the B-splines coefficients with KM, MVMBD and KMPP for `sigma` = 1 are summarized in Table 4.17 and in the boxplots shown in Fig. 4.8. None of the coefficient clustering alternatives prove to be better, in terms of accuracy, than FMBD.

TABLE 4.17.
MODEL 3 SUMMARY STATISTICS FOR COEFFICIENT CLUSTERING, 3-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.328	0.336	0.02929	24460	4	0.1718
	Mean	0.3282	0.339	0.03587	24460	3.646	0.1751
	Variance	0.001158	0.001178	0.000899	59580	0.5632	0.0003907
MV MBD	Median	0.32	0.336	0.02884	24460	2	0.1885
	Mean	0.3275	0.3379	0.03469	24460	2.479	0.1957
	Variance	0.001204	0.001189	0.000876	57340	0.3699	0.0005232
KMPP	Median	0.328	0.336	0.03261	24460	3	0.1719
	Mean	0.3294	0.3396	0.03596	24460	3.516	0.1782
	Variance	0.001087	0.001072	0.000813	57880	0.4762	0.0004313

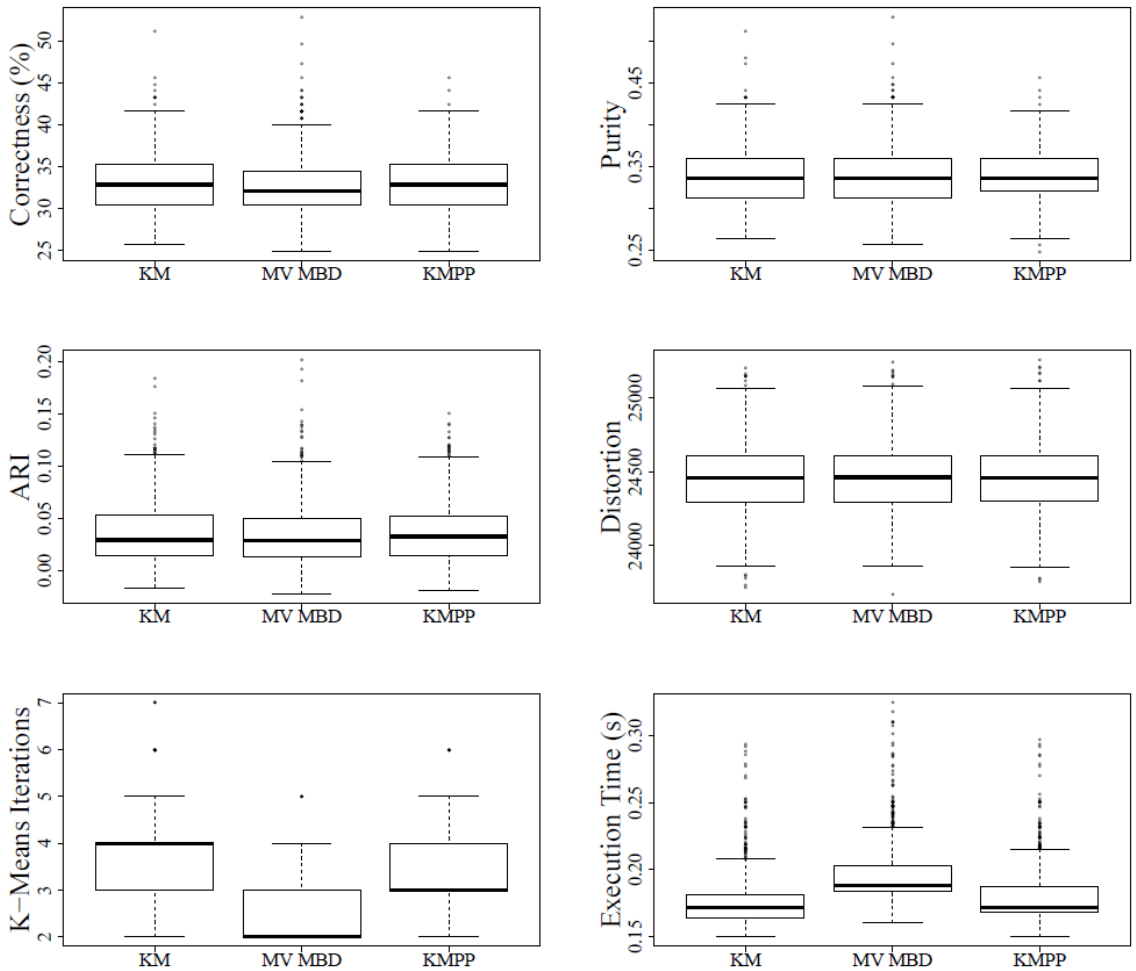


Fig. 4.8. Model 3, 3-way distribution of CoPADIT measures for `sigma` = 1

4.3.3 Missing Data

The results obtained by performing clustering on models with 25% missing data missing data for `sigma` = 1 are summarized in Table 4.18 and in the boxplots shown in Fig. 4.9.

TABLE 4.18.
MODEL 3 SUMMARY STATISTICS FOR 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

`Pmiss` = 0.25, `sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.432	0.448	0.2012	20850	4	0.5069
	Mean	0.4363	0.4506	0.203	20850	4.091	0.5238
	Variance	0.001355	0.001115	0.001315	55870	0.6514	0.00385
MV MBD	Median	0.432	0.448	0.2221	20860	4	0.7054
	Mean	0.434	0.4503	0.2235	20860	4.108	0.729
	Variance	0.001202	0.000987	0.001589	56310	0.7811	0.006569
FMBD	Median	0.496	0.52	0.2831	21090	3	0.3721
	Mean	0.5011	0.5199	0.2841	21090	2.674	0.3923
	Variance	0.002354	0.001743	0.001873	56690	0.4702	0.003326
KMPP	Median	0.432	0.448	0.1969	20850	4	0.5142
	Mean	0.4356	0.4503	0.2004	20850	4.093	0.5314
	Variance	0.001282	0.001017	0.00135	55050	0.661	0.003973
FKMPP	Median	0.496	0.512	0.2671	21100	3	0.1938
	Mean	0.4952	0.5165	0.2706	21100	3.507	0.2087
	Variance	0.002498	0.001818	0.002036	57120	0.5245	0.002053

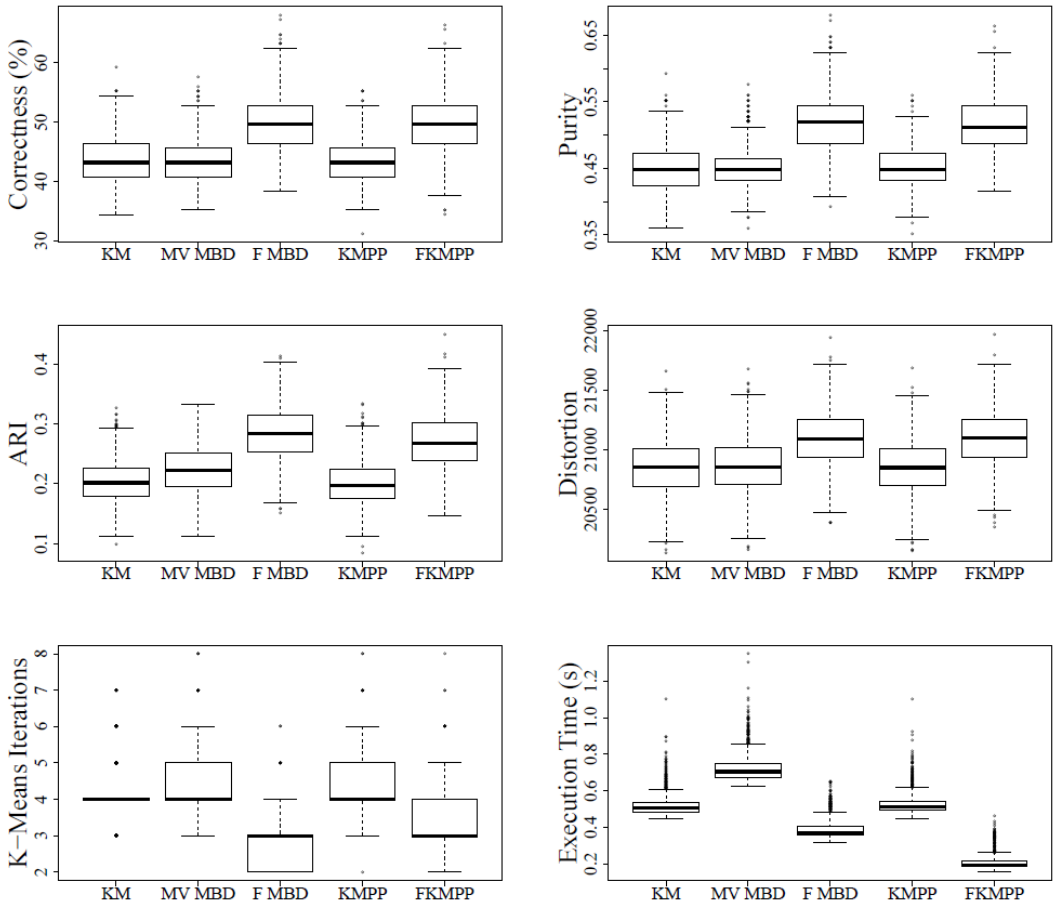


Fig. 4.9. Model 3, 5-way distribution of CoPADIT measures for `sigma` = 1 with 25% missing values.

4.4 Model Four

Table 4.19 shows the parameters that provide the best results for the function approximation for clustering purposes in different noise scenarios.

TABLE 4.19.
MODEL 4 OPTIMAL PARAMETERS.

Parameter	Value
Intercept	*True*
Degree of polynomial	*3*
Degrees of freedom	*4*
Oversampling Factor	*1*

The five-way comparison, the coefficient clustering and missing data results are presented using the values from Table 4.19.

4.4.1 Five-Way Comparison

The results for `sigma` = 1 are summarized in Table 4.20 and in Fig. 4.10.

TABLE 4.19.
MODEL 4 SUMMARY STATISTICS, 5-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.57	0.59	0.307	1894	3	~ 0
	Mean	0.5692	0.5865	0.3071	1894	3.394	3.428e-04
	Variance	0.002879	0.001819	0.003109	3753	0.4652	1.338e-06
MV MBD	Median	0.58	0.6	0.3176	1891	3	0.02082
	Mean	0.5806	0.5955	0.3183	1891	2.667	0.02272
	Variance	0.002751	0.001852	0.003224	3743	0.4646	0.0000837
FMBD	Median	0.63	0.64	0.3737	1935	2	0.0000837
	Mean	0.6293	0.6398	0.3772	1933	2.092	0.1178
	Variance	0.002007	0.001471	0.003132	4007	0.1357	0.0004105
KMPP	Median	0.58	0.59	0.3123	1895	3	1.559e-03
	Mean	0.5719	0.5886	0.3089	1894	3.364	2.060e-03
	Variance	0.003266	0.001969	0.003619	3695	0.512	7.644e-06
FKMPP	Median	0.62	0.64	0.3657	1940	3	0.09325
	Mean	0.611	0.6347	0.3643	1938	2.902	0.09774
	Variance	0.003055	0.001581	0.003792	3934	0.4789	0.0003549

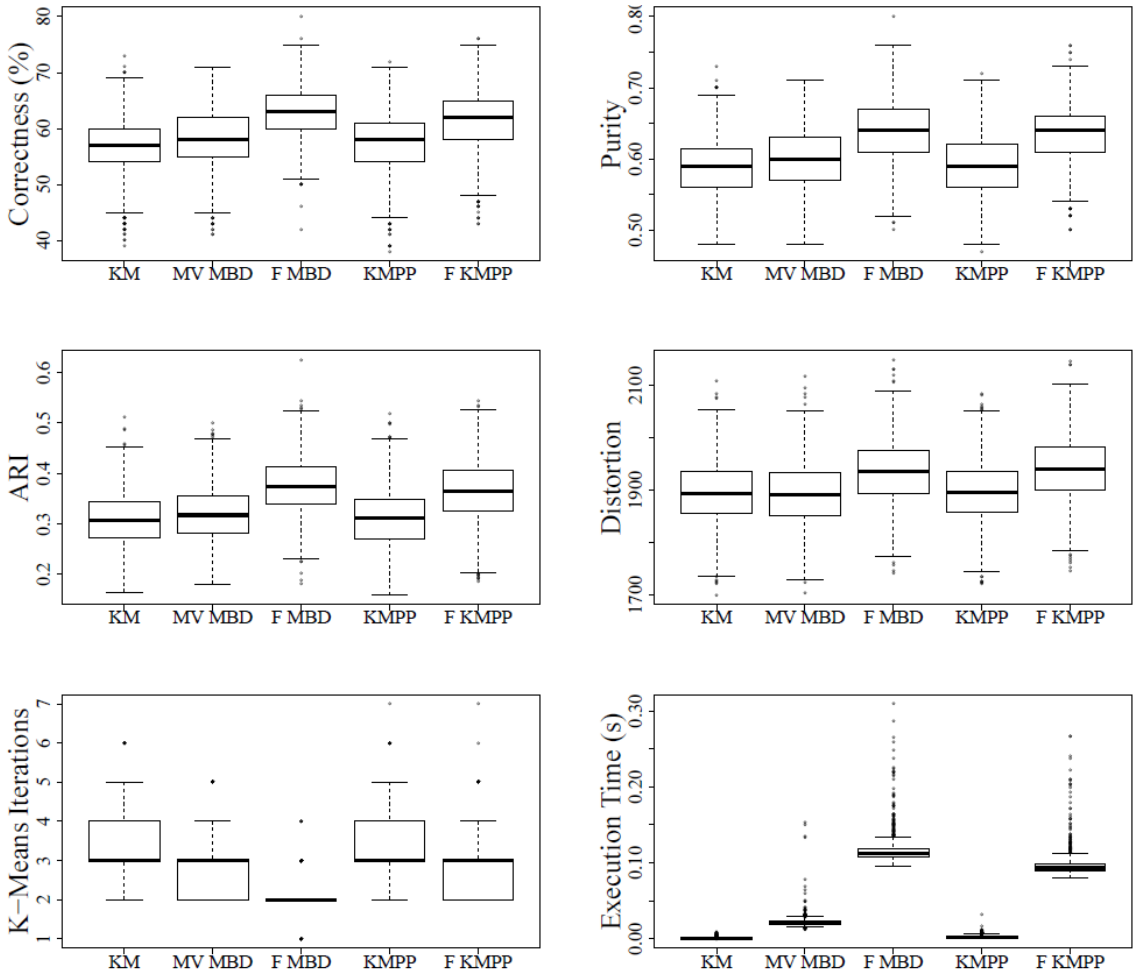


Fig. 4.10. Model 4, 5-way distribution of CoPADIT measures for `sigma` = 1

The p-values for the paired t-test for equality of means of correctness, purity, and ARI for all methods are collected in Table 4.21.

TABLE 4.21.
MODEL 4 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

`sigma` = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	8.819e-09	7.121e-144	2.180e-01	1.902e-68
	Purity		6.516e-10	9.160e-175	2.055e-01	1.118e-147
	ARI		4.773e-09	5.082e-175	4.058e-01	7.642e-113
MV MBD	Correctness	8.819e-09	-	1.862e-112	8.873e-06	4.561e-42
	Purity	6.516e-10		3.554e-140	1.129e-06	3.960e-113
	ARI	4.773e-09		9.662e-141	1.006e-06	7.727e-83
FMBD	Correctness	7.121e-144	1.862e-112	-	1.140e-127	6.603e-26
	Purity	9.160e-175	3.554e-140		3.311e-156	1.049e-06
	ARI	5.082e-175	9.662e-141		1.585e-155	4.526e-16
KMPP	Correctness	2.180e-01	8.873e-06	1.140e-127	-	9.396e-60
	Purity	2.055e-01	1.129e-06	3.311e-156		5.936e-131
	ARI	4.058e-01	1.006e-06	1.585e-155		3.352e-102
FKMPP	Correctness	1.902e-68	4.561e-42	6.603e-26	9.396e-60	-
	Purity	1.118e-147	3.960e-113	1.049e-06	5.936e-131	
	ARI	7.642e-113	7.727e-83	4.526e-16	3.352e-102	

4.4.2 Coefficient Clustering

The results we obtain after clustering the B-splines coefficients with KM, MVMBD and KMPP for `sigma` = 1 are summarized in Table 4.22 and in the boxplots shown in Fig. 4.11.

TABLE 4.22.
MODEL 4 SUMMARY STATISTICS FOR COEFFICIENT CLUSTERING, 3-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.39	0.4	0.04746	2208	3	0.09115
	Mean	0.3918	0.403	0.05964	2201	2.86	0.09005
	Variance	0.002657	0.002565	0.002801	10170	0.4108	0.0002097
MV MBD	Median	0.38	0.4	0.044	2209	2	0.09656
	Mean	0.3913	0.402	0.05949	2203	2.019	0.1006
	Variance	0.003078	0.002975	0.003077	10550	0.09473	0.0003052
KMPP	Median	0.38	0.39	0.03823	2216	3	0.09169
	Mean	0.3853	0.3959	0.05258	2211	2.718	0.09144
	Variance	0.002736	0.002726	0.002814	9648	0.4069	0.0002386

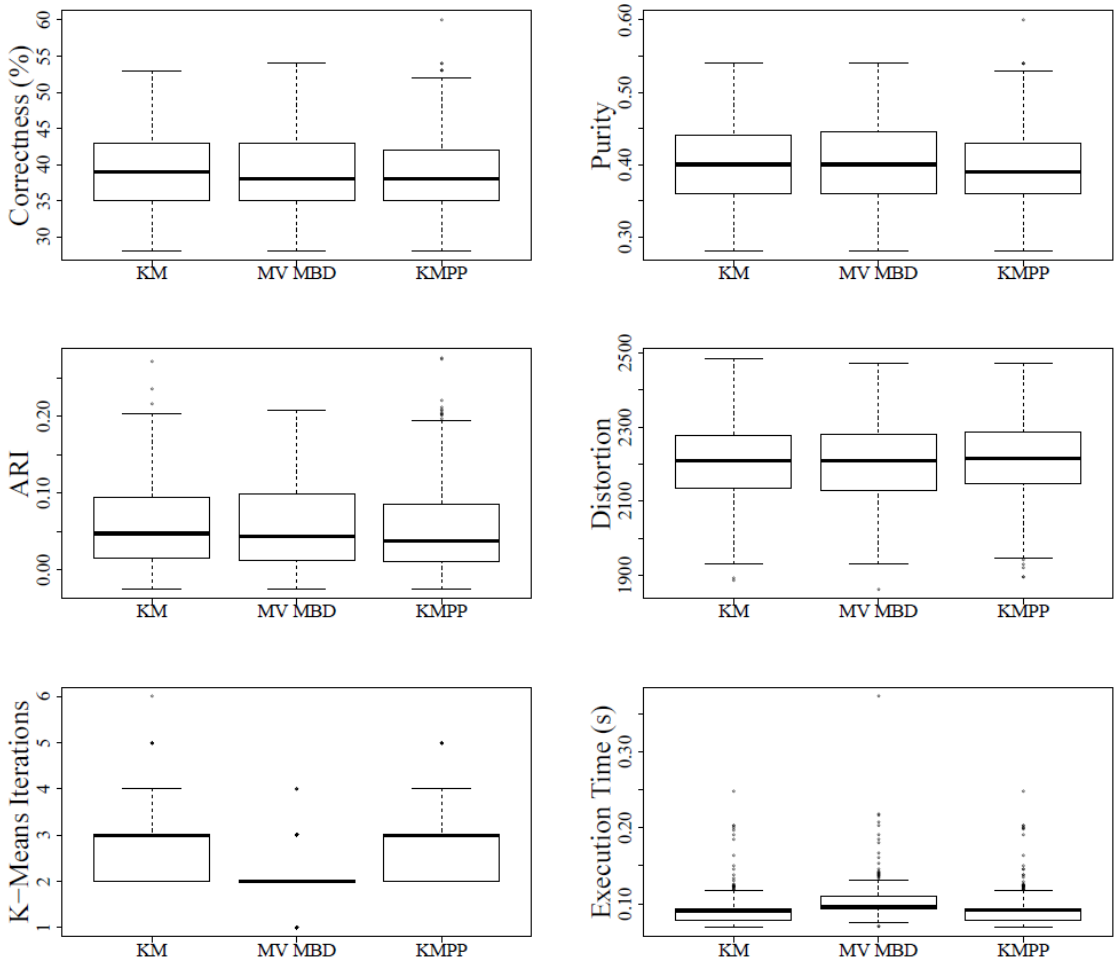


Fig. 4.11. Model 4, 3-way distribution of CoPADIT measures for `sigma` = 1

4.4.3 Missing Data

The results obtained by performing clustering on models with 25% of missing data for `sigma` = 1 are summarized in Table 4.23 and in the boxplots shown in Fig. 4.12.

TABLE 4.23.
MODEL 4 SUMMARY STATISTICS FOR 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

`Pmiss` = 0.25, `sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.57	0.58	0.2858	1653	3	0.04323
	Mean	0.5634	0.5785	0.2843	1654	3.363	0.0405
	Variance	0.002783	0.001874	0.003211	3909	0.4737	0.0001433
MV MBD	Median	0.57	0.58	0.2896	1650	2	0.06248
	Mean	0.5691	0.583	0.2897	1651	2.54	0.0616
	Variance	0.002112	0.001546	0.002901	3832	0.4288	0.0001986
FMBD	Median	0.6	0.61	0.3238	1682	2	0.1094
	Mean	0.5959	0.607	0.3205	1683	2.119	0.1135
	Variance	0.00208	0.001551	0.003212	4030	0.149	0.0003599
KMPP	Median	0.56	0.58	0.2813	1654	3	0.04614
	Mean	0.5626	0.5783	0.2818	1655	3.329	0.04209
	Variance	0.002683	0.001798	0.003144	3844	0.4372	0.0001676
FKMPP	Median	0.58	0.6	0.3082	1687	3	0.09373
	Mean	0.5802	0.5995	0.3085	1688	2.981	0.09381
	Variance	0.003047	0.00187	0.003815	4091	0.4551	0.0002576

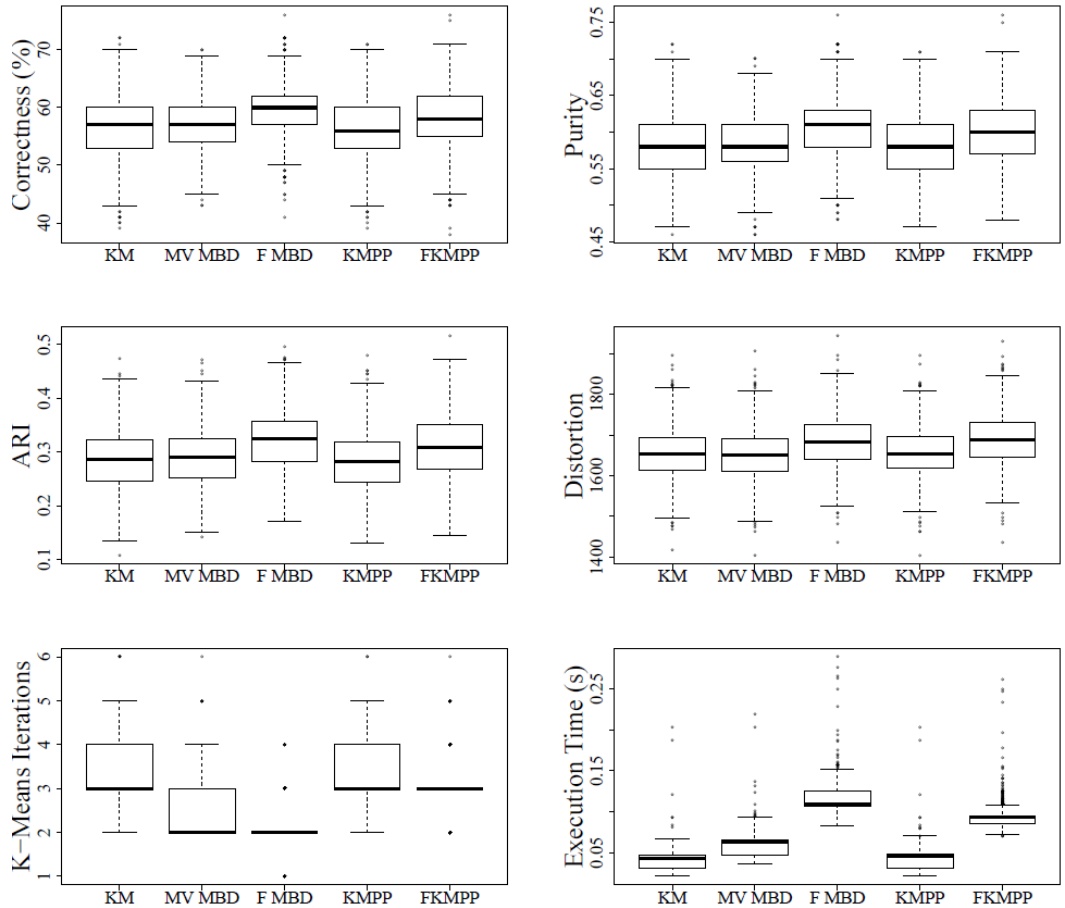


Fig. 4.12. Model 4, 5-way distribution of CoPADIT measures for `sigma` = 1 with 25% missing values.

All these results support the superiority of FMBD over the alternative techniques. A further insight will be given in section 5 (*Conclusions*).

4.5 Real Data

According to the section 3.7 (*Real Data*), climate data is collected for the regions indicated in Table 3.16. These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program [46].

The data of temperature at 2m above the surface and precipitation will be clustered using the values in Table 4.24 for the parameters involved.

TABLE 4.24.
REAL DATA CLUSTERING: PARAMETERS USED.

Parameter	Temperature Data Clustering Value	Precipitation Data Clustering Value
Intercept	$True$	$True$
Degree of polynomial	3	3
Degrees of freedom	20	50
Oversampling Factor	$^{12}/_{365}$	1
Bootstrapping Replicas	25	25

Similarly, to what we did on the simulated models, temperature and precipitation data has been clustered using all the methods we are comparing. Since all the initialization algorithms are non-deterministic, we have run K-Means 1000 times for each of them. The results for the temperature are reported in Table 4.25 and in Figure 4.13, whereas those for the precipitation are summarized in Table 4.26 and Figure 4.15. KM and KMPP are notably worse than the other techniques.

Temperature

TABLE 4.25.
TEMPERATURE SUMMARY STATISTICS, 5-WAY COPADIT.

Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.7055	0.7717	0.6436	1751000	2	0.01561
	Mean	0.7717	0.8213	0.707	1971000	2.345	0.01122
	Variance	0.01703	0.008003	0.01767	3,98E+14	0.2462	8,19E-02
MV MBD	Median	0.9291	0.9291	0.8617	1751000	1	0.6067
	Mean	0.8165	0.8571	0.7632	1686000	1.256	0.6465
	Variance	0.01583	0.006461	0.01211	1,96E+13	0.1907	0.01193
FMBD	Median	0.926	0.926	0.8522	1758000	2	1.238
	Mean	0.9234	0.9243	0.85	1756000	1.7	1.236
	Variance	0.0007094	0.000314	0.000523	267800000	0.2102	0.0001588
KMPP	Median	0.7055	0.7717	0.6436	1751000	2	0.06252
	Mean	0.7761	0.8263	0.7196	1730000	2.062	0.07111
	Variance	0.01363	0.005824	0.01122	6,47E+13	0.06822	0.0004704
FKMPP	Median	0.926	0.926	0.8522	1758000	2	1.222
	Mean	0.8575	0.8793	0.79	1766000	2.104	1.223
	Variance	0.0133	0.00611	0.0108	1,86E+13	0.1173	3,58E-02

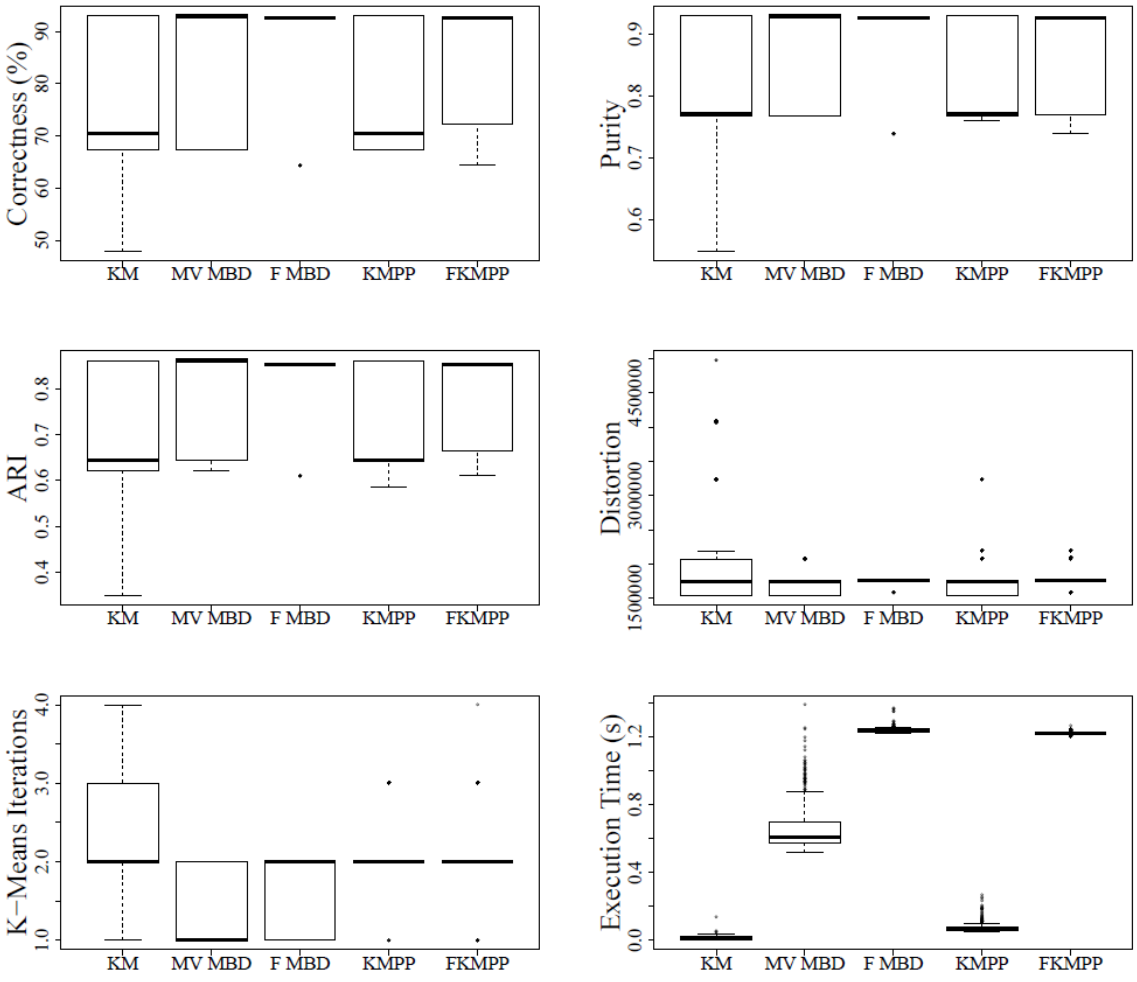


Fig. 4.13. Temperature data, 5-way distribution of CoPADIT measures.

In Fig. 4.14 the bimodal distributions of the accuracy measures for the five methods are displayed. The crosses mark the mean values, while the dots mark the median of the corresponding method's density plot according to the color code shown in the legends. Note that in the case of FMBD there is no real bimodality and thus the corresponding values are consistently better.

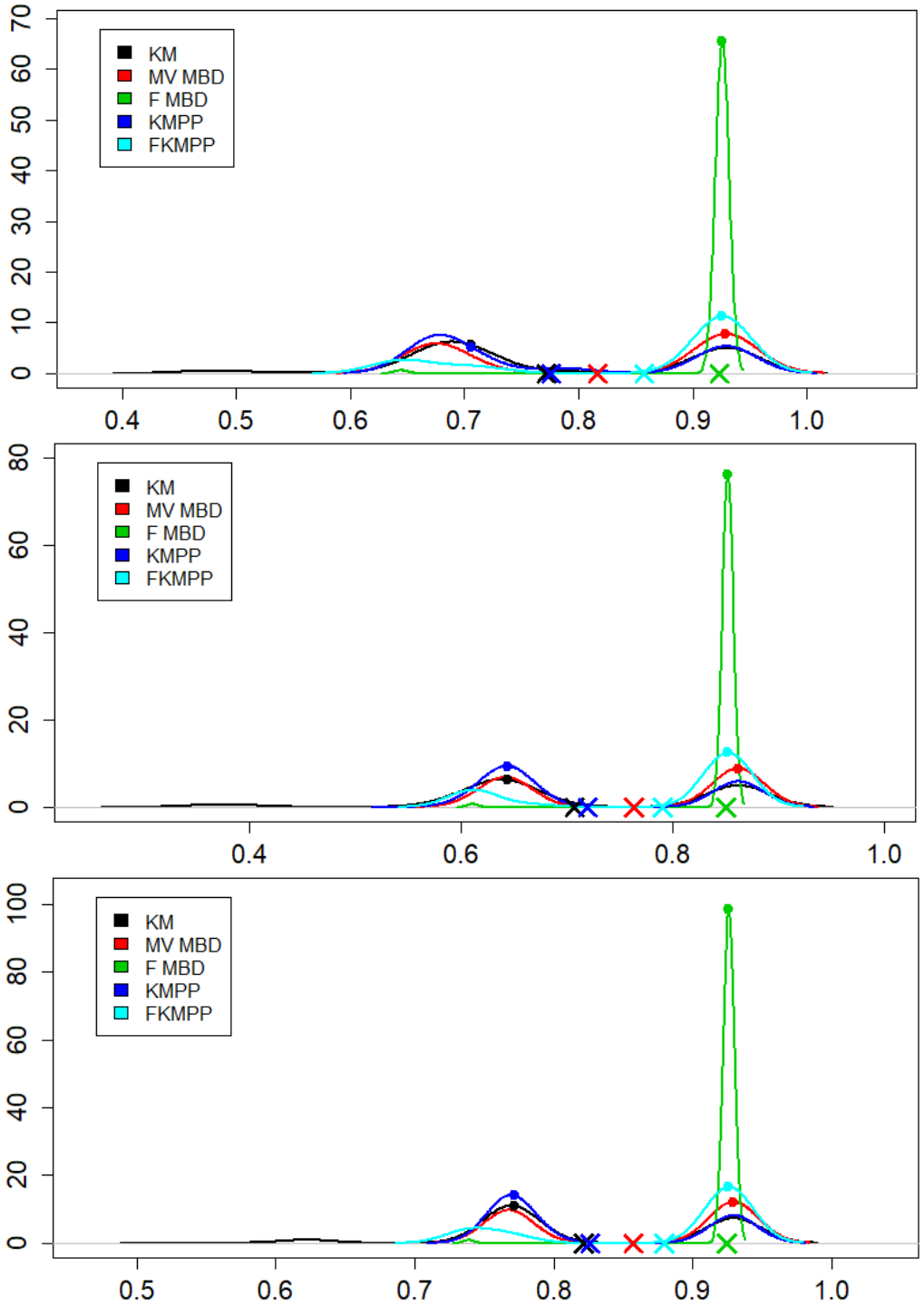


Fig. 4.14. Temperature dataset correctness (top panel), purity (middle panel) and ARI (bottom panel) density plots.

Precipitation

TABLE 4.26.
PRECIPITATION SUMMARY STATISTICS, 5-WAY COPADIT.

Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.5921	0.6378	0.3592	2498000	2	0.01562
	Mean	0.59	0.6364	0.3746	2525000	2.502	0.0119
	Variance	0.0119	0.007367	0.01758	2,78E+13	0.2903	9,57E-02
MV MBD	Median	0.6614	0.685	0.3931	2404000	2	0.6657
	Mean	0.6421	0.6697	0.3866	2419000	1.72	0.6975
	Variance	0.003034	0.001457	0.003402	1,97E+12	0.2078	0.01314
FMBD	Median	0.7039	0.726	0.558	2409000	2	2.891
	Mean	0.7095	0.734	0.5687	2418000	1.631	2.924
	Variance	0.0004446	0.000583	0.001111	9.362E+08	0.2411	0.01082
KMPP	Median	0.5213	0.5906	0.2881	2501000	2	0.06531
	Mean	0.5264	0.5873	0.3046	2510000	2.433	0.07426
	Variance	0.00979	0.005495	0.01005	1,18E+13	0.2758	0.0005386
FKMPP	Median	0.7039	0.726	0.51	2449000	2	2.302
	Mean	0.6628	0.7128	0.5091	2466000	2.548	2.309
	Variance	0.01045	0.005997	0.01545	1,15E+13	0.4261	0.0005511

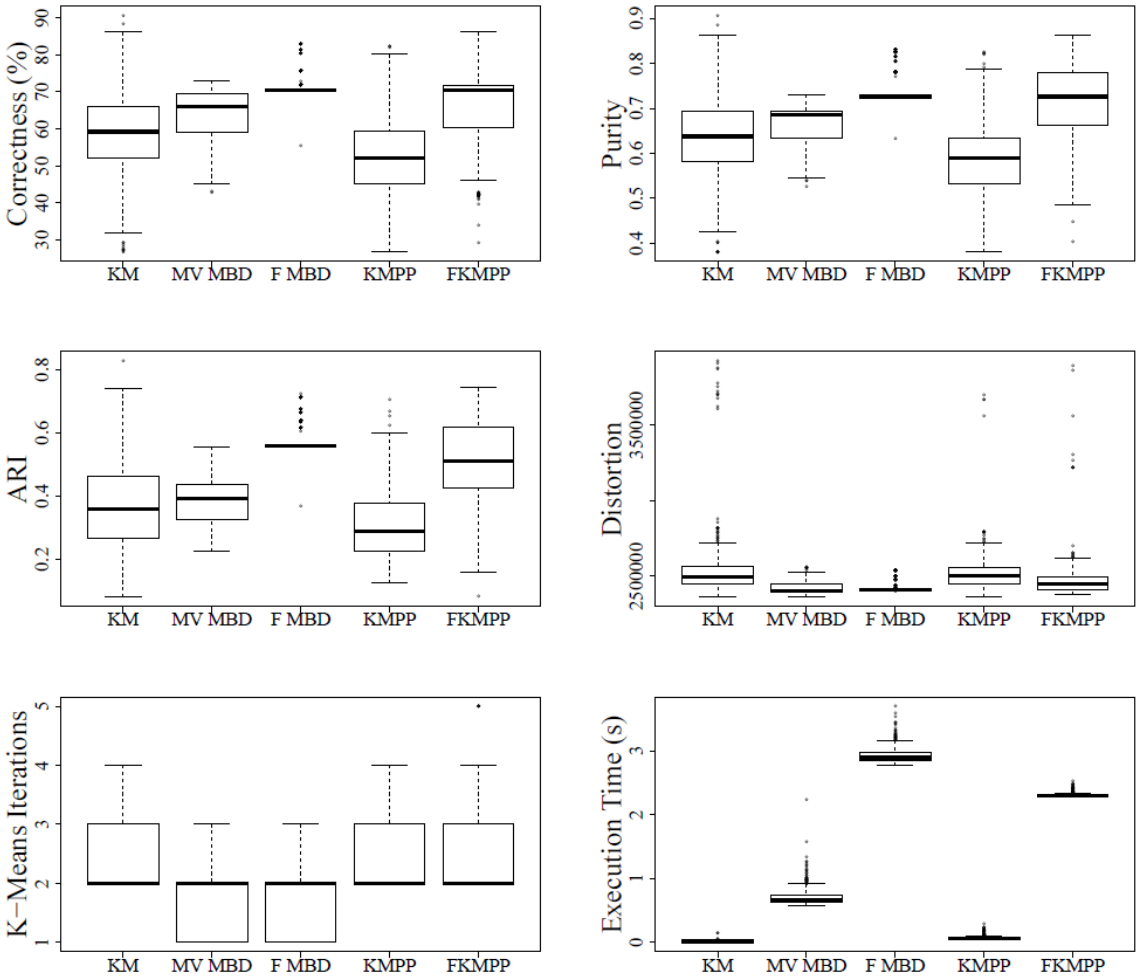


Fig. 4.15. Precipitation data, 5-way distribution of CoPADIT measures.

We see in Fig. 4.16 that the accuracy measures have multi-modal distributions. Now FMBD and FKMPP are revealed as the best options. The crosses show the mean values, while the dots illustrate the median of the corresponding method's density plot.

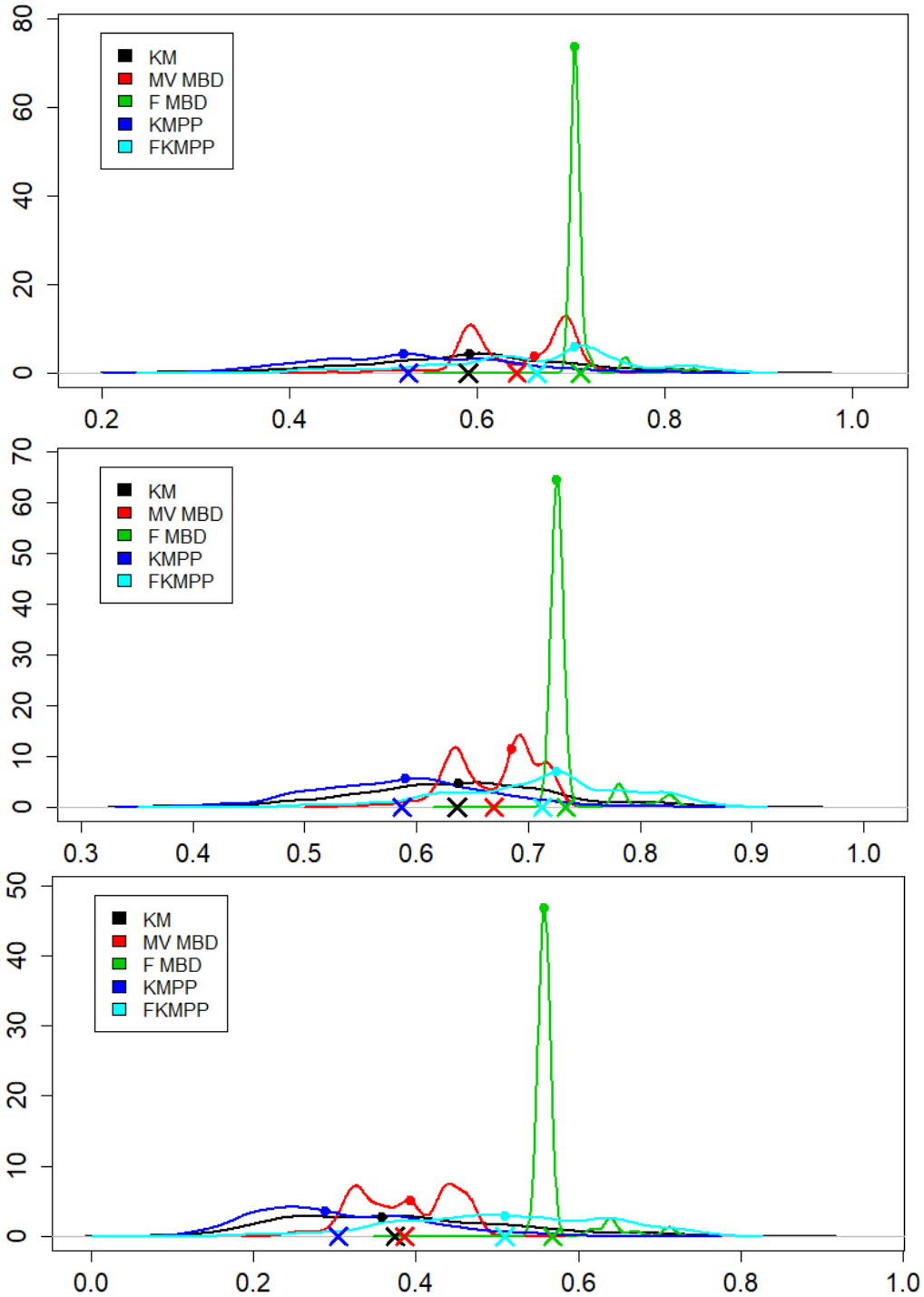


Fig. 4.16. Precipitation dataset correctness (top panel), purity (middle panel) and ARI (bottom panel) density plots.

4.6 Qualitative Summary

Finally, a descriptive summary of the performance of FMBD with respect to the alternatives is provided in Table 4.27. The table collects the situations in which the method proposed outperforms the rest (↑), it is as good as the best alternative (=) or some other method provides better results (↓).

TABLE 4.27.
QUALITATIVE SUMMARY OF THE MEDIAN / MEAN /VARIANCE STATISTICS OF FMBD'S PERFORMANCE FOR THE COPADIT MEASURES

		Measure					
		Correctness	Purity	ARI	Distortion	Iterations	Time
Data clustering	Model 1	↑/↑/↑	↑/↑/=	↑/↑/=	↓/↓/=	↑/↑/↑	↓/↓/↓
	Model 2	=/↑/↑	=/↑/↑	↑/↑/↑	↓/↓/↑	↑/↑/↑	↓/↓/↓
	Model 3	↑/↑/↓	↑/↑/↓	↑/↑/↓	↓/↓/↓	↑/↑/↑	↓/↓/↓
	Model 4	↑/↑/↑	=/↑/=	↑/↑/=	↓/↓/↓	↑/↑/↑	↓/↓/↓
Coefficient clustering	Model 1	↑/↑/↑	↑/↑/↑	↑/↑/↑	↑/↑/↑	↑/↑/↑	↓/↓/=
	Model 2	↑/↑/↑	↑/↑/↑	↑/↑/↑	↑/↑/↑	↑/↑/↑	↓/↓/=
	Model 3	↑/↑/↓	↑/↑/↓	↑/↑/↓	↑/↑/↑	↓/↓/↓	↓/↓/↓
	Model 4	↑/↑/=	↑/↑/↑	↑/↑/=	↑/↑/↑	=/=/↓	↓/↓/=
Missing data	Model 1	↑/↑/=	↑/↑/=	=/↑/=	↓/↓/↓	↑/↑/↑	↓/↓/=
	Model 2	↓/↓/↓	↓/↓/=	↓/↓/↑	↓/↓/↓	=/↑/↓	↓/↓/↓
	Model 3	=/=/=	↑/=/=	=/↑/=	↓/↓/↓	↑/↑/↑	↓/↓/↓
	Model 4	↑/=/=	↑/=/=	↑/↑/=	↓/↓/↓	=/↑/↑	↓/↓/↓
Real datasets	Temperature data	=/↑/↑	=/↑/↑	↓/↑/↑	=/=/↑	=/↓/↓	↓/↓/↑
	Precipitation data	=/↑/↑	=/↑/↑	↑/↑/↑	=/↑/↑	=/↑/↓	↓/↓/↓

5. CONCLUSIONS

In the previous section and in the Appendix, the clustering evaluation measures have been used to evaluate quantitatively how good the outputs for the different methods are. What do all these tables and figures mean? Is MBD-based initialization of K-Means for data approximated by B-Splines (FMBD) reliable for clustering?

Following the reasoning established in the Goals section (subsection 1.4), the three main points of the project are outlined and explained.

1. **Addressing the transformation of input multivariate data into functions for the clustering analysis.**

The introductory page for all models and real data (subsections 4.1 to 4.5) has a summary of the OSF and DF parameters tested. The experimentation results performed to find these parameters are collected in the Appendix.

First of all, we can see that the higher the DF, the tighter the fit of the curve to the data. Noisy data (i.e. for high values of `sigma`) require more smoothing and a looser fit, and hence low values of DF.

$$\uparrow Noise \Rightarrow \downarrow DF\ required$$

Secondly, it can be seen that an increase in the OSF does not provide better clustering results. That is, transforming input multivariate data into a function to obtain more observations does not offer an advantage.

Alternatively, as proven with the temperature data clustering (section 4.5), lowering the OSF means that a better clustering output is obtained. This is because the samples are placed at relevant intervals, and other secondary information is discarded. These *x*-axis values are enough to perform the clustering and, in some occasions, improve the results. Note that the factor is set to $^{12}/_{365}$, which represents one sample per month.

The relationship between the clustering evaluation measures used and the OSF is hard to materialize and no certain conclusion can be reached.

$$\uparrow OSF \not\Rightarrow \uparrow CoPADIT$$

Drawing these two points together, DF and OSF interact together in order to provide new observations that are more suitable for clustering than the original data. The methods that rely on the B-Spline curve fitting, FMBD and FKMPP, provide better correctness, purity and ARI measures than the other methods in all models except for model 2, which requires higher DF in order for the methods to be comparable.

Moreover, the distortion of the functional methods tends to be slightly higher than the other methods because the centroids for each cluster are found with the fitted points and not the original ones, and the sum of squared distances to the points inside a cluster (i.e. the distortion) is a bit higher.

From the execution time perspective, the conversion of the original multivariate data into functional data makes these methods slower, but the number of iterations needed for convergence is generally lower in FMBD and FKMPP. The curve fitting takes longer for larger datasets, which can be a major concern in the context of *big data*.

All in all, in general it can be said that:

$$FDA \Rightarrow \uparrow accuracy\ and \uparrow computational\ time$$

Hence, the transformation of input multivariate data into functions (FDA) for a clustering analysis comes down to a tradeoff between the increase in accuracy and the increase in computational time.

2. Assessing the clustering output results using a set of performance evaluation measures and comparing the proposed method to other initialization methods.

Generally speaking, FMBD clustering works well with models as well as real data. It is an advantageous solution that offers higher accuracy than other techniques at the cost of a longer computational time and a slightly higher distortion. A qualitative summary of the tables and plots is found in Table 4.6.

The method proposed has higher correctness, purity and ARI indices for almost all noise levels. It can be seen that every t-test pairwise comparison suggests that the hypothesis of equality of means has to be rejected. In particular, all comparisons that involve FMBD and FKMPP yield drastically small p-values.

In the case of the five-way comparison for model 1, `sigma` = 1, seen in Table 4.4, the correctness median is 10% higher than the second-best method (FKMPP), and the correctness mean is more than 8% higher. In most cases, the initialization makes K-Means converge in two iterations.

B-Spline coefficient clustering does not provide almost any advantages with respect to the other methods tested, except for a faster convergence of the K-Means algorithm that does not translate into lower execution time.

In the cases in which there are missing data, FMBD is still a reliable method, only failing when clusters are less distinguishable and functions present high oscillations (model 2).

Furthermore, in the case of real datasets of temperature and precipitation, functional MBD-based initialization rises as a reliable way of clustering giving consistently better results as proven by the low variance.

3. Determining whether MBD-based initialization is an advantageous solution for K-Means clustering in the case of functional data.

FMBD has proven to be a consistent and robust method for K-Means initialization, hence being an advantageous solution for clustering in the case of functional data, working both for modeled situations and real-world scenarios.

6. SOCIAL AND ECONOMIC IMPACT

6.1 Social and Economic Implications of the Project

The attraction of data analysis field in the telecommunications sector has skyrocketed. Companies demand an increasing number of professionals that possess knowledge in the areas of data science and machine learning [48]. One of the tools used by the data scientists is clustering [49].

The estimated value of the data economy in Europe represented 1.87% of the GDP of the member countries in 2015 (272,000 million euros) and is expected to reach 4.7% in 2020. In addition, 65% of companies run the risk of becoming irrelevant or uncompetitive if they do not adopt Big Data. The data analysis market in Spain grows 30% each year and employed 10,500 professionals in 2015 [50].

Improving the outcome of a well-known unsupervised classification algorithm like K-Means leads to enhanced results in various fields. A fascinating research area related to clustering is biomedicine. The expansion in sophisticated techniques for data capturing has made the use of data analysis for solving health related-problems possible [1], [7]. It is pleasant and rewarding to think that producing a more precise clustering output entails better disease treatment and ultimately more lives being saved.

Furthermore, clustering is not only used in this sector, but also others like market analysis. Segmentation and customer targeting of a company depend heavily on these kind of techniques [51], and therefore improved clustering results provide increased profit.

The research method proposed in this Final Year Project, FMBD, pushes the limits of K-Means' accuracy and emphasizes the role of Universidad Carlos III de Madrid as an outstanding research institution in the matter of statistics.

6.2 Relationship with Telecom Engineering

Following the lines of thought proposed in subsection 6.1, it is possible to apply customer segmentation to the telecommunications industry in the same way that is done in market analysis. This comes with the advantages of improving quality of service for users and enhanced targeted marketing for enterprises [8].

In addition to being a relevant algorithm in the data science field, K-Means clustering is related to Communication Theory. It relies on data classification according to distance, in the same way as received symbols in an AWGN channel are matched to an element of the constellation. Voronoi regions define the sections of space that are used for these decisions. The symbols of a constellation can be thought of centroids of a cluster in this analogy.

6.3 Budget

Table 6.1 summarizes the costs for this project. A total of 590 hours have been invested in its completion. The 510 hours dedicated to this project by the undergraduate researcher include 500 hours from a Research Collaboration with University Departments Scholarship from the Spanish Ministry of Education. The workload distribution has been done according to Fig. 6.1.

Additionally, it must be noted that the hardware budget has been computed according to the specifications in Table 4.2.

TABLE 6.1.
SUMMARY OF PROJECT COSTS.

Costs Summary			
Item			Total Cost (€)
Concept	Hours	Cost per hour	
Undergraduate Research Scholarship [52]	500	-	2,000
Undergraduate Researcher (extra hours)	10	15	150
Project Manager	80	45	3,600
Hardware* [53]	-	-	1,549 · 17% = 263
Software**	-	-	0
Total			6,013

* The cost of the hardware for the project is 1,549, with corresponding amortization of 25% [54] for 8 months.
** The Open Source License of R-Studio Desktop, as well as Microsoft Office for students, can be downloaded for free [55], [56].

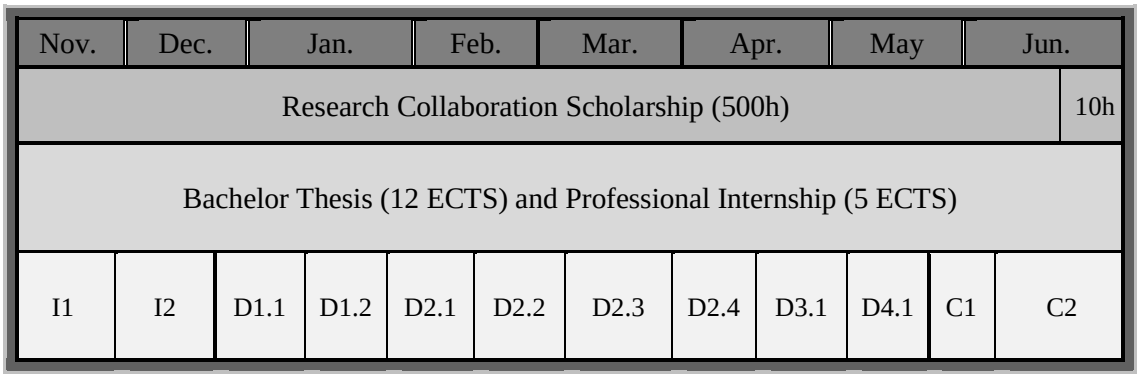


Fig. 6.1. Undergraduate researcher project timeline according to Table 2.1.

7. DISCUSION AND FUTURE STUDIES

We finally consider the limitations we have encountered in our study as well as some related future research lines.

7.1 Limitations

FMBD works best in high noise scenarios with clearly separate clusters made of functions with low oscillations. The main constraints of MBD-based initialization of K-Means for functional data are:

1. Higher distortion values due to the smoothing of the datapoints used to initialize the algorithm.
2. Greater computational complexity and longer execution time, especially requiring the tuning of the degrees of freedom parameter involved in the functional approximation of the input data.

The degrees of freedom can be seen as a smoothing parameter for noisy data. High-frequency noise is filtered out by the function approximation procedure. In the signal processing world, this is known as a *low pass filter*. Setting this parameter to a low value means that we have a smoother resulting curve, and hence being equivalent to a filter with a lower cutoff frequency.

In addition, it can be seen from the *OSF and DF Behavior According to Input Data* section of the Appendix that higher OSF values do not provide an advantage for a more accurate clustering result. This is the reasoning behind only testing the DF's for model 3 and not the OSF influence on the results, as seen in the Appendix. Model 4 resembles model 1 in the sense that it does not have high oscillations of the functions used to generate the dataset, and hence the same parameters are used for the function approximation for both models.

From another point of view, the bootstrapping factor's influence on the results has not been exhaustively documented in the project, but is also a relevant consideration for the results. A higher bootstrapping factor, B, is expected to produce better correctness, purity and ARI for FMBD and MVMBD, though a higher computational cost. To this respect, it is important to take into account that a B = 5 is occasionally enough for FMBD to yield better results than the other proposed methods.

7.2 Proposed Method as an R Package

The coding written to implement FMBD has been organized in scripts which can be easily documented and crafted into an R Package. It is of the author's intention to make it publicly available for download to nurture future research in the area. From a personal perspective, it would be gratifying to see a real impact of this project in the community.

7.3 Future Research Lines

Although this project provides a thorough insight to the use of MBD for functional data clustering with K-Means, there are still further improvements that can be done for tighter curve fitting, and predictably better clustering results. *Time warping* is proposed as a solution to situations in which a compression or dilation of the *x*-axis would help in clustering. There are some R packages available online that can be explored to implement this modification to the curve-fitting process [57].

8. REFERENCES

[1] H. Sørensen, J. Goldsmith, L. M. Sangalli, “An Introduction with Medical Applications to Functional Data Analysis,” Statistics in Medicine, vol 32, No. 30, pp. 5222-5240, September 2013.

[2] P. Hall, D. S. Poskitt, and B. Presnell, “A Functional Data-Analytic Approach to Signal Discrimination,” Technometrics, vol. 43, No. 1, pp. 1-9, February 2001.

[3] H. E. Jones, and N. Bayley, “The Berkeley Growth Study,” Child Development, vol. 12, No. 2, pp. 167-173, June 1941.

[4] D. Soni, “Supervised vs. Unsupervised Learning,” Towards Data Science. [Online]. Available at https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d [Accessed June 5, 2019].

[5] Wikipedia, “Cluster Analysis,” Wikipedia. [Online]. Available at: https://en.wikipedia.org/wiki/Cluster_analysis. [Accessed June 5, 2019].

[6] R. Henao et al., “Patient Clustering with Uncoded Text in Electronic Medical Records,” US National Library of Medicine AMIA Annual Symposium Proceedings Archive, vol. November 2013.

[7] M. Barahona and P. Matthews, “Clustering disease progression through feature landscapes and clusters of symptoms,” EPSRC Centre for Mathematics of precision Healthcare. [Online]. Available at http://www.imperial.ac.uk/mathematics-precision-healthcare/research/clustering-symptoms-in-disease-progression-/ [Accessed June 5, 2019].

[8] L. Ye, C. Qiuru, X. Haixu, L. Yijun and Z. Guangping, “Customer Segmentation for Telecom with the K-Means Clustering Method”, Information Technology Journal, vol. 12, No. 3, pp. 409-413, 2013.

[9] J. C. Barca and G. Rumantir, "A Modified K-means Algorithm for Noise Reduction in Optical Motion Capture Data," 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007), Melbourne, Qld., pp. 118-122, July 2007.

[10] M. E. Celebi, H. A. Kingravi, and P. A. Vela, “A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm,” Expert Systems with Applications, vol. 40, No. 1, pp. 200-210, September 2012. Cornell University Archive, Ithaca, New York.

[11] Z. Zhenjie, B. T. Dai, and A. Tung. “On the Lower Bound of Local Optimums in K-Means Algorithm,” in Proceedings of the 6th IEEE International Conference on Data Mining, December 2006. Hong Kong.

[12] S. López-Pintado and J. Romo, “On the Concept of Depth for Functional Data,” Journal of the American Statistical Association, vol. 104, No. 486, pp. 718-734, June 2009. [Online]. Available at: https://www.jstor.org/stable/40592217 [Accessed June 5, 2019].

[13] A. Torrente and J. Romo, “Initializing k-means clustering by bootstrap and data depth,” submitted to *Journal of Classification* (under revision).

[14] European Patent Office, “Guidelines for Examination, 3.3 – Mathematical Methods”, *European Patent Office*. [Online]. Available at: https://www.epo.org/law-practice/legal-texts/html/guidelines2018/e/g_ii_3_3.htm [Accessed June 11, 2019].

[15] European Parliament and of the Council, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance),” European Parliament and of the Council, OJ L 119, 4.5.2016, p. 1–88, April 27, 2016. [Online]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 [Accessed June 3, 2019].

[16] K. Willems, “Choosing R or Python for Data Analysis? An infographic,” *Data Camp*. [Online]. Available at https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis [Accessed June 5, 2019].

[17] MIT Libraries, “Statistical Software: R,” *MIT Libraries*. [Online]. Available at https://libguides.mit.edu/stat/r [Accessed June 5, 2019].

[18] Georgia Tech Library, “Introduction to R Studio,” *Georgia Tech Library*. [Online]. Available at https://libguides.gatech.edu/c.php?g=890639&p=6403717 [Accessed June 5, 2019]

[19] H. Nyquist, "Certain Topics in Telegraph Transmission Theory," Transactions of the American Institute of Electrical Engineers**,** vol. 47, No. 2, pp. 617-644, April 1928.

[20] J. Ding, V. Tarokh, and Y. Yang, “Model Selection Techniques – An Overview,” *IEEE Signal Processing Magazine*, October 2018. Cornell University Archive, Ithaca, New York.

[21] A. V. Oppenheim, A. S. Willsky and S. J. Nawab, *Signals and Systems*, Second Edition. Edinburgh Gate, Harlow, UK: Pearson, 2014.

[22] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Second Edition. Berlin, Germany: Springer Series in Statistics, 2010.

[23] H. J. Landau, "Sampling, data transmission, and the Nyquist rate," Proceedings of the IEEE, vol. 55, No. 10, pp. 1701-1706, October 1967.

[24] M. Ç. Pinar, “Overdetermined Systems of Linear Equations,” *Encyclopedia of Optimization*, pp. 1925-1928, January 2001.

[25] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Second Edition. Hoboken, New Jersey, US: John Wiley & Sons, 2001.

[26] H. Chenchouni, T. Menasria, S. Neffar, S. Chafaa, L. Bradai, R. Chaibi, M. N. Mekahlia and D. Bendjoudi, "Spatiotemporal diversity, structure and trophic guilds of insect assemblages in semi-arid Sabkha ecosystem," *PeerJ*, 3:e860, March, 2015.

[27] J. McGonagle, G. Pilling,V. Tembo, et al., "Gaussian Mixture Model," *Brilliant.org* [Online]. Available at https://brilliant.org/wiki/gaussian-mixture-model/ [Accessed June 5, 2019].

[28] R. Misra, "Inference using EM algorithm," *Towards Data Science*. [Online]. Available at https://towardsdatascience.com/inference-using-em-algorithm-d71cccb647bc [Accessed June 5, 2019].

[29] R. Gandhi, "K-Means Clustering – Introduction to Machine Learning Algorithms," *Towards Data Science*. [Online]. Available at https://towardsdatascience.com/k-means-clustering-introduction-to-machine-learning-algorithms-c96bf0d5d57a [Accessed June 5, 2019].

[30] L. Kaufman, P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", *Wiley Online Library*, March 8 1990.

[31] M. Maechler, "Partitioning Around Medoids", *RDocumentation*, Available at https://www.rdocumentation.org/packages/cluster/versions/2.0.7-1/topics/pam [Accessed June 5, 2019].

[32] Wikipédia, "Algorithme des k-médoïdes," *Wikipédia*. [Online]. Available at: https://fr.wikipedia.org/wiki/Algorithme_des_k-m%C3%A9do%C3%AFdes#/ [Accessed June 5, 2019].

[33] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding", *Stanford University InfoLab*, June 2006.

[34] C. D. Manning, P. Raghavan and H. Schütze, "Evaluation of Clustering," in *Introduction to Information Retrieval*, First Edition. Cambridge, England. Cambridge University Press, April 2009.

[35] L. Hubert and P. Arabie. "Comparing Partitions," *Journal of Classification*, vol. 2, No. 1, pp. 193-218, February 1985.

[36] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, No. 1, pp. 1-26, January 1979.

[37] J. Fox, "Bootstrapping Regression Models Appendix to An R and S-PLUS Companion to Applied Regression," *Stanford School of Humanities and Sciences, Department of Statistics*, January 2002.

[38] S. Glen, "Bootstrap Sample: Definition, Example," *Statistics How To*. [Online]. Available at https://www.statisticshowto.datasciencecentral.com/bootstrap-sample/ [Accessed June 5, 2019].

[39] A. Canty and B. Ripely, “Package ‘boot’”, *The Comprehensive R Archive Network*, April, 2019. Available at: https://cran.r-project.org/web/packages/boot/boot.pdf [Accessed June 11, 2019].

[40] I. Cascos, A. López and J. Romo, “Data Depth in Multivariate Statistics”, *Boletín de Estadística e Investigación Operativa*, vol. 27, No. 3, pp. 151-174, October 2011.

[41] S. Ribecca, “Parallel Coordinates Plot,” *The Data Visualisation Catalogue*. [Online]. Available at: https://datavizcatalogue.com/methods/parallel_coordinates.html [Accessed June 5, 2019]

[42] Q. Chaudhari, “Additive White Gaussian Noise (AWGN),” *Wireless Pi*, [Online]. Available at: https://wirelesspi.com/additive-white-gaussian-noise-awgn/ [Accessed June 5, 2019]

[43] A. Artés Rodríguez F. Pérez González, J. Cid Sueiro, R. López Valcarce, C. Mosquera Nartallo, F. Pérez Cruz, “Modulación y detección en canales gaussianos”, in *Comunicaciones Digitales*. Pearson (discontinued), 2012.

[44] A. Leroy, A. Marc, O. Dupas, J. L, Rey, “Functional Data Analysis in Sport Science: Example of Swimmers’ Progression Curves Clustering,” *Applied Sciences*, vol. 8, No. 10. September, 2018.

[45] Q. Li, R. Li, K. Ji and W. Dai, "Kalman Filter and Its Application," *8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS)*, p74-77, Tianjin, China, November 2015.

[46] National Aeronautics and Space Administration, POWER Project Data Sets. [Online]. Available at: https://power.larc.nasa.gov/ [Accessed June 5, 2019]

[47] M. Kottek, J. Grieser, C. Beck, B. Rudolf and F. Rubel, “World Map of the Köppen-Geiger Climate Classification Updated”, *Meteorologische Zeitschrift*, vol. 15, No. 3, pp. 259-263, June 2006.

[48] K. Pattabiraman, “The Most Promising Jobs of 2019,” *LikedIn Blog*, January 10, 2019. [Online]. Available at: https://blog.linkedin.com/2019/january/10/linkedins-most-promising-jobs-of-2019 [Accessed June 7, 2019]

[49] G. Seif, “The 5 Clustering Algorithms Data Scientists Need to Know,” *Towards Data Science*, February 5, 2018. [Online]. Available at: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 [Accessed June 7, 2019]

[50] Fundación Cotec para la Innovación, “Generación de Talento Big Data en España,” *Cotec*, 2017. [Online]. Available at: http://informecotec.es/media/BIG-DATA-FINAL-web.pdf [Accessed June 11, 2019].

[51] T. Evgeniou, “Cluster Analysis and Segmentation,” *INSEAD Analytics Couse Sessions*, May, 2019. [Online]. Available at: https://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html [Accessed June 7, 2019]

[52] Ministerio de Educación y Formación Profesional, “Becas de colaboración,” *Ministerio de Educación y Formación Profesional.* November 7, 2018. [Online]. Available at: http://www.culturaydeporte.gob.es/servicios-al-ciudadano-mecd/gl/catalogo/general/educacion/998142/ficha.html [Accessed June 11, 2019].

[53] ASUSTeK Computer Inc., “ASUSPRO B,” *ASUS Shop*. [Online]. Available at: https://eshop.asus.com/es-ES/portatiles/serie-profesional/asuspro-b [Accessed June 11, 2019].

[54] Agencia Tributaria, “Tabla de coeficientes de amortización lineal,” *Agencia Tributaria.* [Online]. Available at: https://www.agenciatributaria.es/AEAT.internet/Inicio/_Segmentos_/Empresas_y_profesionales/Empresas/Impuesto_sobre_Sociedades/Periodos_impositivos_a_partir_de_1_1_2015/Base_imponible/Amortizacion/Tabla_de_coeficientes_de_amortizacion_lineal_.shtml [Accessed June 11, 2019].

[55] "Download RStudio - RStudio", *RStudio*, 2019. [Online]. Available at: https://www.rstudio.com/products/rstudio/download. [Accessed June 7, 2019]

[56] Microsoft, “Office 365 Education,” *Microsoft*. [Online]. Available at: https://www.microsoft.com/en-us/education/products/office. [Accessed June 7, 2019]

[57] T. Giorgino, *Package ‘dtw’*, May 18, 2018. Available at: https://cran.r-project.org/web/packages/dtw/dtw.pdf [Accessed June 7, 2019].

APPENDIX

We assemble here, for completeness, the results obtained for other values of the parameters, as described in the main text of this manuscript, as well as the study carried out to find the optimal values of the parameters, organized in the following way:

Pages III to L:

5-Way Comparison for `sigma` = 0, 0.5, 1, 1.5 and 2 for all models with the corresponding p-values of the t-test.

Coefficient clustering for `sigma` = 1 and 2 for all models.

5-Way Comparison for 25%, 50% and 75% missing values for `sigma` = 1 for all models.

Pages LI to LXVIII:

`OSF` and `DF` behavior according to input data.

NOTE: The values of Purity, ARI, Distortion and Iterations (PADI) given in this section correspond to those of the FMBD method. From these values we can obtain the optimal parameters included in the *Results* section (Table 4.3, Table 4.9, Table 4.14 and Table 4.19) of the main text. Additionally, the optimal parameters of FKMPP coincide with those of FMBD, which are not included for simplicity.

The following values of `sigma` are considered:

- Model 1: `sigma` = 0.5, 1, 1.5, 2 and 10.
- Model 2: `sigma` = 1, 1.5, 2 and 10.
- Model 3: `sigma` = 0.5, 1, 1.5, 2 and 10.
- Model 4: see model 1 due to their similarity.

The tables found in the main section are repeated in the Appendix so that the reader does not have to jump backwards and forwards to compare the desired values.

Model One – Five-Way Comparison

TABLE A.1.
SUMMARY STATISTICS FOR MODEL 1, 5-WAY COPADIT FOR SIGMA = 0.

sigma = 0							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	1	1	1	0	1	~ 0
	Mean	1	1	1	0	1	9.581e-04
	Variance	0	0	0	0	0	1.338e-05
MV MBD	Median	1	1	1	0	1	0.07811
	Mean	1	1	1	0	1	0.0889
	Variance	0	0	0	0	0	0.0006279
FMBD	Median	1	1	1	0	1	0.1762
	Mean	1	1	1	0	1	0.1982
	Variance	0	0	0	0	0	0.002319
KMPP	Median	1	1	1	0	1	~ 0
	Mean	1	1	1	0	1	4.158e-03
	Variance	0	0	0	0	0	5.658e-05
FKMPP	Median	1	1	1	0	1	0.1003
	Mean	1	1	1	0	1	0.11250
	Variance	0	0	0	0	0	0.00111

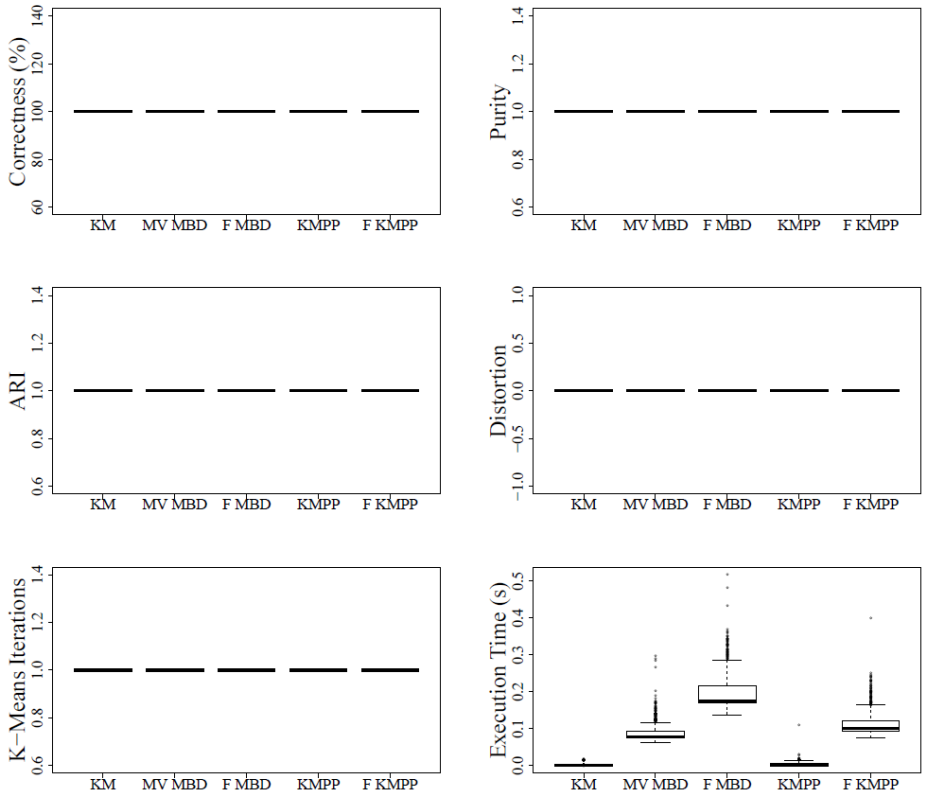


Fig. A.1. Model 1, 5-way CoPADIT measures distribution for sigma = 0.

TABLE A.2.
SUMMARY STATISTICS FOR MODEL 1, 5-WAY COPADIT FOR SIGMA = 0.5.

`sigma` = 0.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.83	0.83	0.6936	2429	3	~ 0
	Mean	0.7946	0. 8387	0.7305	2432	2.582	6.254e-04
	Variance	0.02313	0. 01115	0.02002	1963	0.2855	8.667e-06
MV MBD	Median	0.87	0.87	0.7361	2428	2	0.0781
	Mean	0.8261	0. 852	0.7552	2427	2.231	0.07864
	Variance	0.01542	0.008536	0.01292	1372	0.2418	0.000324
FMBD	Median	0.98	0.98	0.9472	2420	1	0.1736
	Mean	0.9764	0.9764	0.9395	2420	1.433	0.1825
	Variance	0.000238	0.000238	0.001416	1288	0.2458	0.001145
KMPP	Median	0.84	0.84	0.7068	2429	3	~ 0
	Mean	0.8005	0.84300	0.7369	2430	2.566	3.295e-03
	Variance	0.02205	0.01051	0.01857	1899	0.2859	3.813e-05
FKMPP	Median	0.97	0.97000	0.9225	2432	2	0.1021
	Mean	0.8666	0. 89920	0.8327	2432	2.157	0.1067
	Variance	0.02428	0.01211	0.02409	1723	0.1685	0.0006205

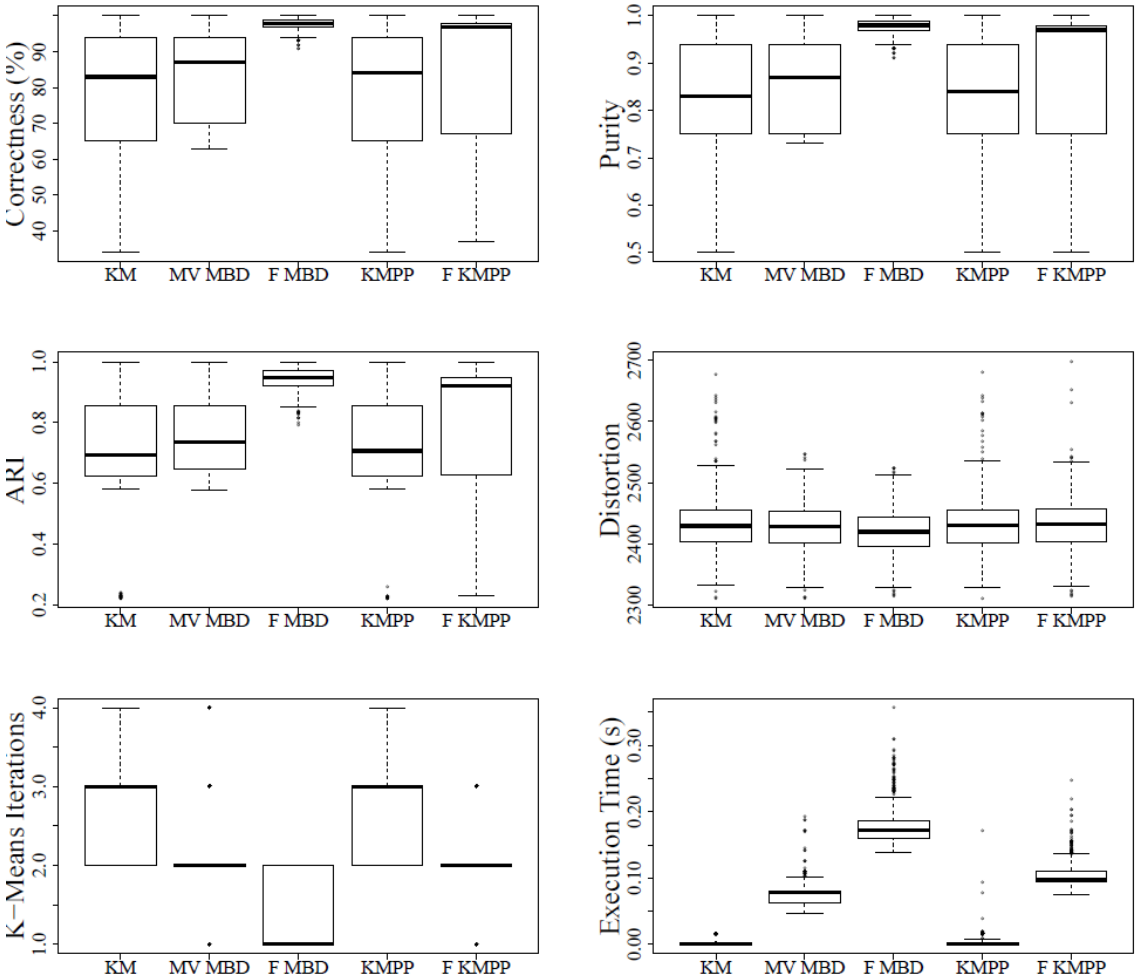


Fig. A.2. Model 1, 5-way CoPADIT measures distribution for `sigma` = 0.5.

TABLE A.3.
SUMMARY STATISTICS FOR MODEL 1, 5-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.66	0.67	0.4294	9568	4	~ 0
	Mean	0.6524	0.667	0.42	9572	3.721	1.032e-03
	Variance	0.005021	0.003507	0.007326	20130	0.5938	6.777e-06
MV MBD	Median	0.67	0.68	0.4369	9575	3	0.08278
	Mean	0.6596	0.6718	0.4315	9573	3.484	0.08855
	Variance	0.004463	0.003272	0.007274	20290	0.6684	0.0004869
FMBD	Median	0.84	0.84	0.6513	9650	2	0.1862
	Mean	0.8253	0.8277	0.6467	9651	1.949	0.1955
	Variance	0.003671	0.002979	0.005546	20170	0.1125	0.001486
KMPP	Median	0.65	0.67	0.4265	9684	4	2.992e-03
	Mean	0.6487	0.6675	0.4193	9687	3.760	3.993e-03
	Variance	0.00466	0.003513	0.007457	20780	0.563	1.947e-05
FKMPP	Median	0.74	0.8	0.6103	9656	3	0.1089
	Mean	0.7404	0.795	0.6099	9659	2.764	0.114
	Variance	0.01134	0.004935	0.008153	20390	0.4387	0.0006833

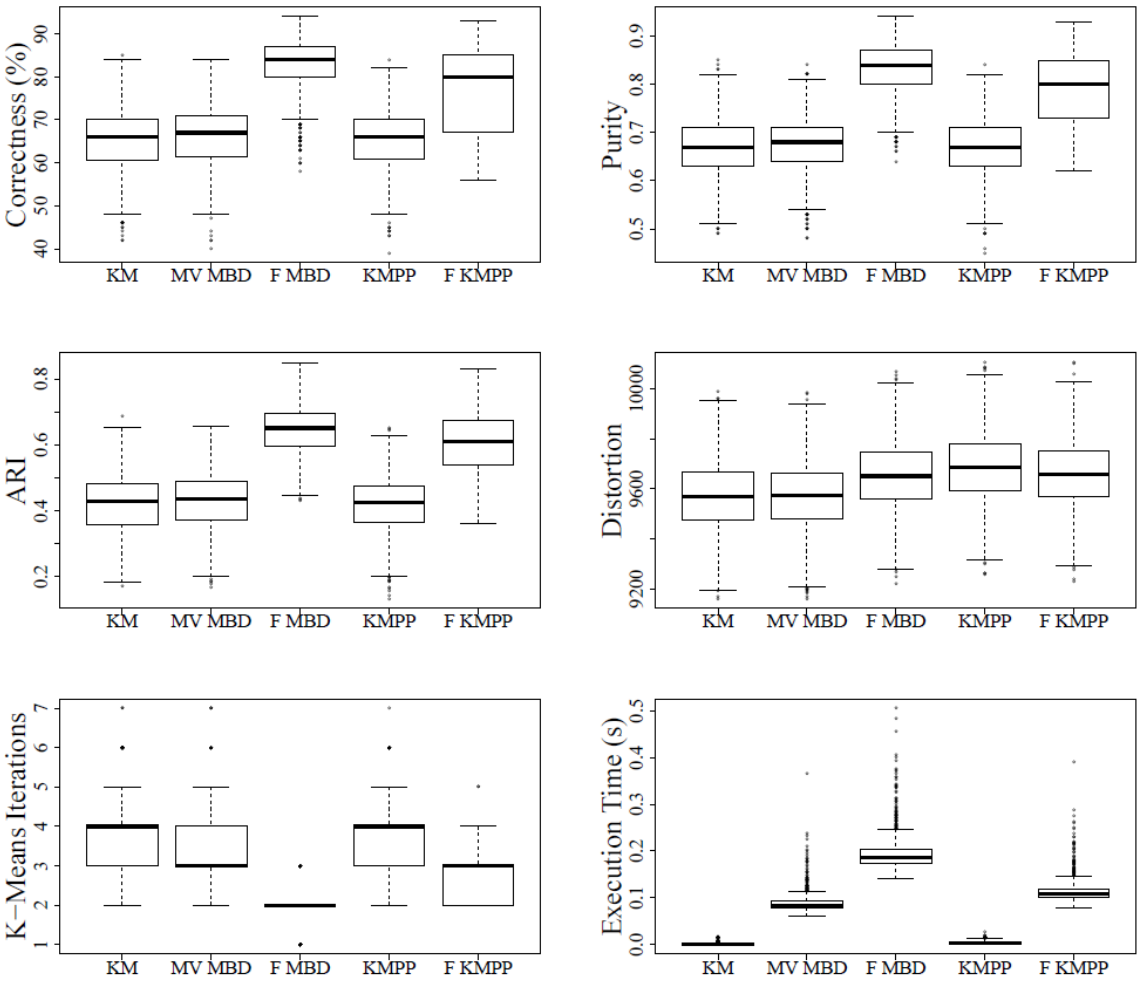


Fig. A.3. Model 1, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.4.
SUMMARY STATISTICS FOR MODEL 1, 5-WAY COPADIT FOR SIGMA = 1.5.

sigma = 1.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.45	0.46	0.1104	21290	4	~ 0
	Mean	0.4513	0.4632	0.1167	21290	3.9760	1.205e-03
	Variance	0.003513	0.003081	0.003687	100600	0.6801	2.883e-05
MV MBD	Median	0.46	0.47	0.1234	21300	4	0.08158
	Mean	0.4576	0.4704	0.1277	21290	3.9970	0.08573
	Variance	0.00364	0.003191	0.003978	102700	0.7577	0.0003143
FMBD	Median	0.67	0.67	0.3834	21620	2	0.1746
	Mean	0.6638	0.6715	0.3858	21630	2.0870	0.1824
	Variance	0.00425	0.003124	0.004948	101400	0.1456	0.0007477
KMPP	Median	0.45	0.46	0.1137	21290	4	2.068e-03
	Mean	0.452500	0.4644	0.1178	21800	3.9620	3.996e-03
	Variance	0.003358	0.003065	0.003507	105200	0.7093	3.593e-05
FKMPP	Median	0.63	0.65	0.369	21640	3	0.1014
	Mean	0.6358	0.659	0.3703	21640	3.1250	0.1051
	Variance	0.005473	0.002921	0.004809	101100	0.3858	0.0003938

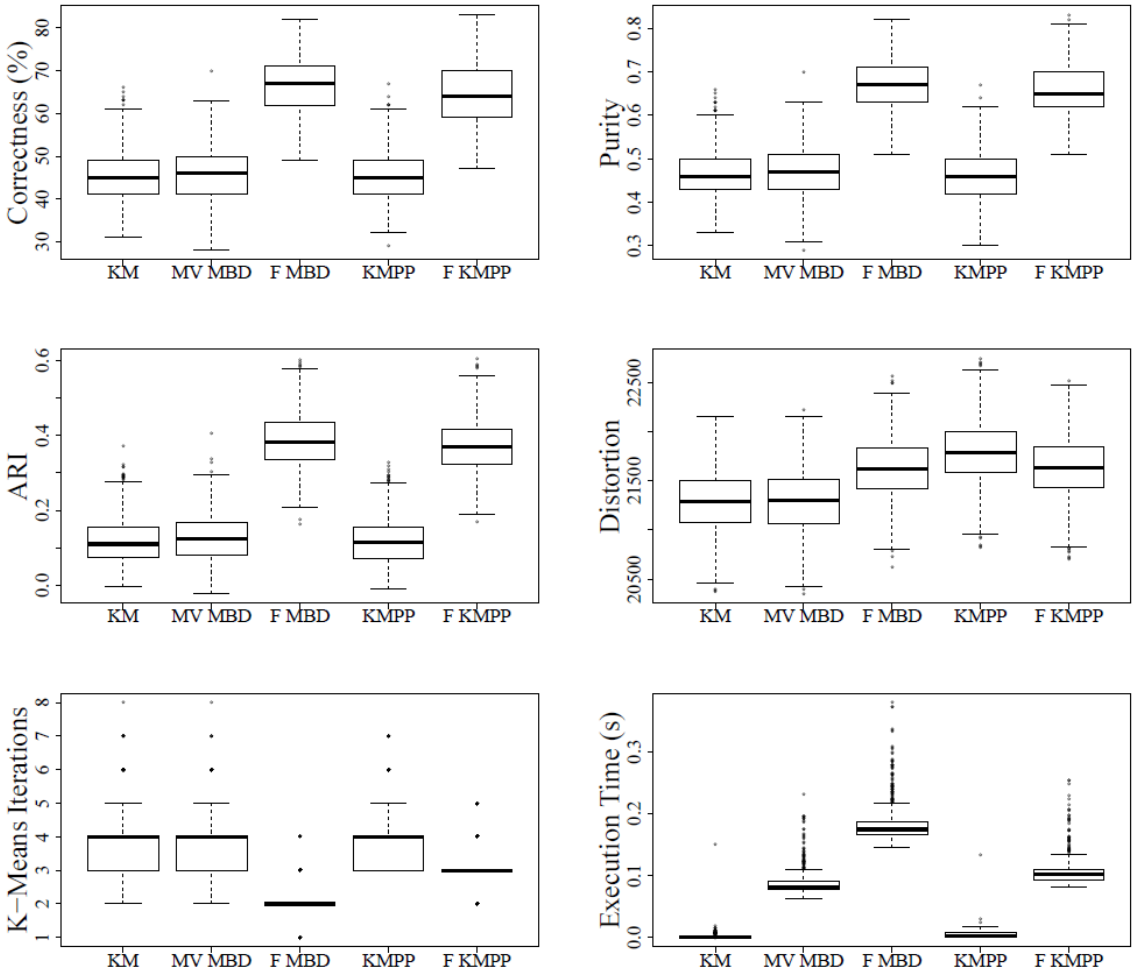


Fig. A.4. Model 1, 5-way CoPADIT measures distribution for sigma = 1.5.

TABLE A.5.
SUMMARY STATISTICS FOR MODEL 1, 5-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.38	0.39	0.03471	37540	4	~ 0
	Mean	0.3811	0.3914	0.03952	37520	3.963	1.072e-03
	Variance	0.001564	0.001515	0.001049	310900	0.6843	3.604e-05
MV MBD	Median	0.38	0.39	0.03532	37540	4	0.07811
	Mean	0.3812	0.3911	0.04065	37540	4.097	0.08228
	Variance	0.001709	0.001656	0.001136	314300	0.8304	0.0002932
FMBD	Median	0.56	0.56	0.2263	38350	2	0.1719
	Mean	0.5569	0.5672	0.2299	38350	2.158	0.1788
	Variance	0.003583	0.002679	0.003713	319700	0.1692	0.0006477
KMPP	Median	0.38	0.39	0.0358	37530	4	~ 0
	Mean	0.3817	0.3923	0.04061	37530	3.981	2.675e-03
	Variance	0.00173	0.001604	0.001088	316300	0.7594	3.285e-05
FKMPP	Median	0.54	0.56	0.2183	38350	3	0.0963
	Mean	0.5437	0.5565	0.2184	38360	3.272	0.1028
	Variance	0.00386	0.002861	0.00358	319200	0.4244	0.000396

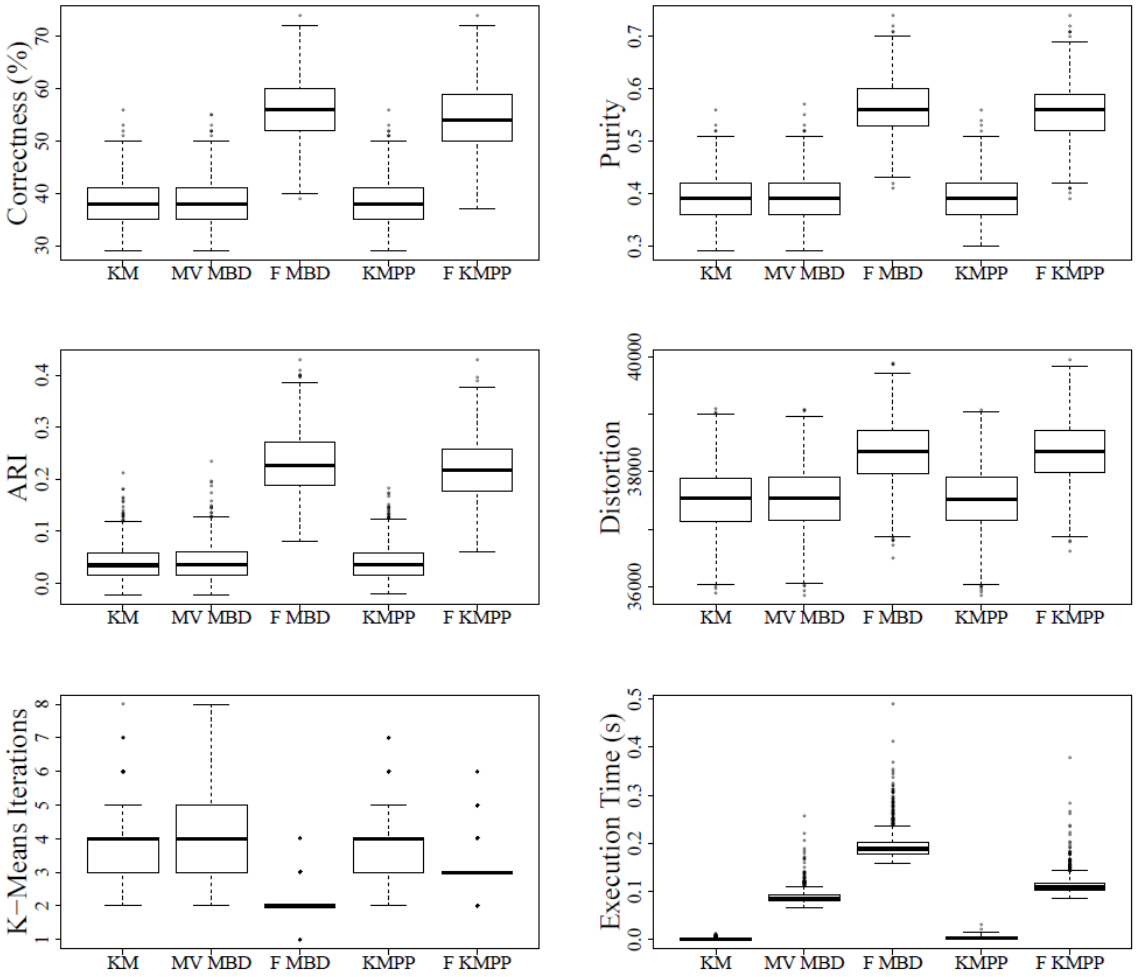


Fig. A.5. Model 1, 5-way CoPADIT measures distribution for sigma = 2.

The p-values for the paired t-test of correctness, purity and ARI for all methods are collected in the following tables for the different values of `sigma`.

TABLE A.6.
MODEL 1 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 0.5.

`sigma` = 0.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	3.285e-07	2.438e-195	3.807e-01	8.839e-25
	Purity		2.035e-03	2.343e-218	3.503e-01	1.101e-34
	ARI		6.718e-06	6.881e-252	2.899e-01	2.066e-50
MV MBD	Correctness	3.285e-07	-	1.231e-197	2.664e-05	2.022e-10
	Purity	2.035e-03		3.622e-226	3.482e-02	4.874e-24
	ARI	6.718e-06		1.267e-275	6.183e-04	1.041e-34
FMBD	Correctness	2.438e-195	1.231e-197	-	5.897e-193	3.176e-90
	Purity	2.343e-218	3.622e-226		4.926e-218	5.360e-90
	ARI	6.881e-252	1.267e-275		6.971e-254	2.377e-88
KMPP	Correctness	3.807e-01	2.664e-05	5.897e-193	-	2.955e-21
	Purity	3.503e-01	3.482e-02	4.926e-218		6.869e-30
	ARI	2.899e-01	6.183e-04	6.971e-254		4.619e-44
FKMPP	Correctness	8.839e-25	2.022e-10	3.176e-90	2.955e-21	-
	Purity	1.101e-34	4.874e-24	5.360e-90	6.869e-30	
	ARI	2.066e-50	1.041e-34	2.377e-88	4.619e-44	

TABLE A.7.
MODEL 1 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

`sigma` = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	5.470e-03	~ 0	9.105e-01	2.290e-160
	Purity		1.763e-02	~ 0	8.227e-01	6.015e-256
	ARI		3.219e-05	~ 0	8.196e-01	1.333e-293
MV MBD	Correctness	5.470e-03	-	~ 0	7.693e-03	4.447e-144
	Purity	1.763e-02		~ 0	3.246e-02	1.084e-242
	ARI	3.219e-05		~ 0	1.211e-05	7.620e-274
FMBD	Correctness	~ 0	~ 0	-	~ 0	4.483e-50
	Purity	~ 0	~ 0		~ 0	1.639e-41
	ARI	~ 0	~ 0		~ 0	8.155e-40
KMPP	Correctness	9.105e-01	7.693e-03	~ 0	-	1.793e-158
	Purity	8.227e-01	3.246e-02	~ 0		6.214e-252
	ARI	8.196e-01	1.211e-05	~ 0		9.524e-294
FKMPP	Correctness	2.290e-160	4.447e-144	4.483e-50	1.793e-158	-
	Purity	6.015e-256	1.084e-242	1.639e-41	6.214e-252	
	ARI	1.333e-293	7.620e-274	8.155e-40	9.524e-294	

TABLE A.8.
MODEL 1 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.5.

sigma = 1.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	0.004167	~ 0	0.7034	~ 0
	Purity		0.0002483	~ 0	0.5418	~ 0
	ARI		1.503e-07	~ 0	6.105e-01	~ 0
MV MBD	Correctness	0.004167	-	~ 0	0.01203	~ 0
	Purity	0.0002483		~ 0	0.0024380	~ 0
	ARI	1.503e-07		~ 0	2.647e-06	~ 0
FMBD	Correctness	~ 0	~ 0	-	~ 0	6.123e-19
	Purity	~ 0	~ 0		~ 0	1.12e-13
	ARI	~ 0	~ 0		~ 0	3.575e-17
KMPP	Correctness	0.7034	0.01203	~ 0	-	~ 0
	Purity	0.5418	0.0024380	~ 0		~ 0
	ARI	6.105e-01	2.647e-06	~ 0		~ 0
FKMPP	Correctness	~ 0	~ 0	6.123e-19	~ 0	-
	Purity	~ 0	~ 0	1.12e-13	~ 0	
	ARI	~ 0	~ 0	3.575e-17	~ 0	

TABLE A.9.
MODEL 1 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 2.

sigma = 2						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	0.9236	~ 0	0.5562	~ 0
	Purity		0.8488	~ 0	0.551	~ 0
	ARI		0.3520	~ 0	0.3608	~ 0
MV MBD	Correctness	0.9236	-	~ 0	0.6188	~ 0
	Purity	0.8488		~ 0	0.4322	~ 0
	ARI	0.3520		~ 0	0.9683	~ 0
FMBD	Correctness	~ 0	~ 0	-	~ 0	4.331e-16
	Purity	~ 0	~ 0		~ 0	1.365e-12
	ARI	~ 0	~ 0		~ 0	1.251e-13
KMPP	Correctness	0.5562	0.6188	~ 0	-	~ 0
	Purity	0.551	0.4322	~ 0		~ 0
	ARI	0.3608	0.9683	~ 0		~ 0
FKMPP	Correctness	~ 0	~ 0	4.331e-16	~ 0	-
	Purity	~ 0	~ 0	1.365e-12	~ 0	
	ARI	~ 0	~ 0	1.251e-13	~ 0	

Model One - Coefficient Clustering

TABLE A.10.
SUMMARY STATISTICS FOR MODEL 1, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.55	0.57	0.2589	9821	3	0.1017
	Mean	0.548	0.5724	0.2662	9828	2.715	0.1109
	Variance	0.006042	0.00366	0.006465	23860	0.4302	0.001383
MV MBD	Median	0.55	0.57	0.2582	9824	2	0.1131
	Mean	0.5451	0.5695	0.2637	9828	1.949	0.1227
	Variance	0.005907	0.003722	0.006438	23970	0.1045	0.001688
KMPP	Median	0.54	0.56	0.2415	9833	2	0.1043
	Mean	0.5358	0.5637	0.255	9834	2.492	0.1131
	Variance	0.006122	0.003627	0.006269	23550	0.3863	0.001452

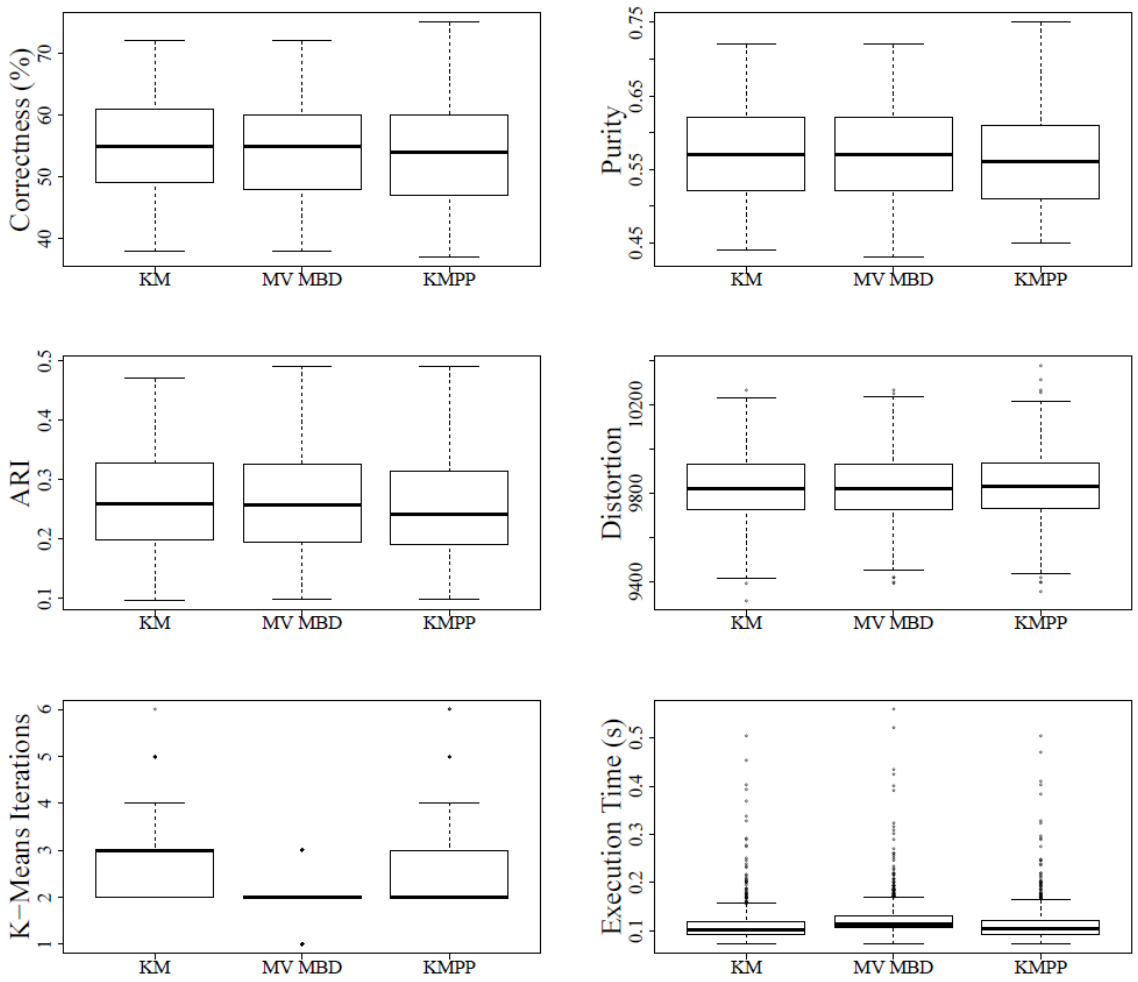


Fig. A.6. Model 1, 3-way CoPADIT measures distribution for sigma = 1.

TABLE A.11.
SUMMARY STATISTICS FOR MODEL 1, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.41	0.43	0.07598	38710	3	0.1029
	Mean	0.4189	0.4378	0.08122	38730	2.711	0.109
	Variance	0.001861	0.001441	0.001304	344200	0.3959	0.0007707
MV MBD	Median	0.41	0.43	0.07491	38730	2	0.1152
	Mean	0.4153	0.4352	0.07915	38740	1.935	0.1201
	Variance	0.001765	0.001365	0.00126	340600	0.09287	0.0008679
KMPP	Median	0.41	0.43	0.0759	38720	2	0.1067
	Mean	0.4162	0.4362	0.08023	38730	2.445	0.1108
	Variance	0.001788	0.001385	0.001267	336600	0.3353	0.0007829

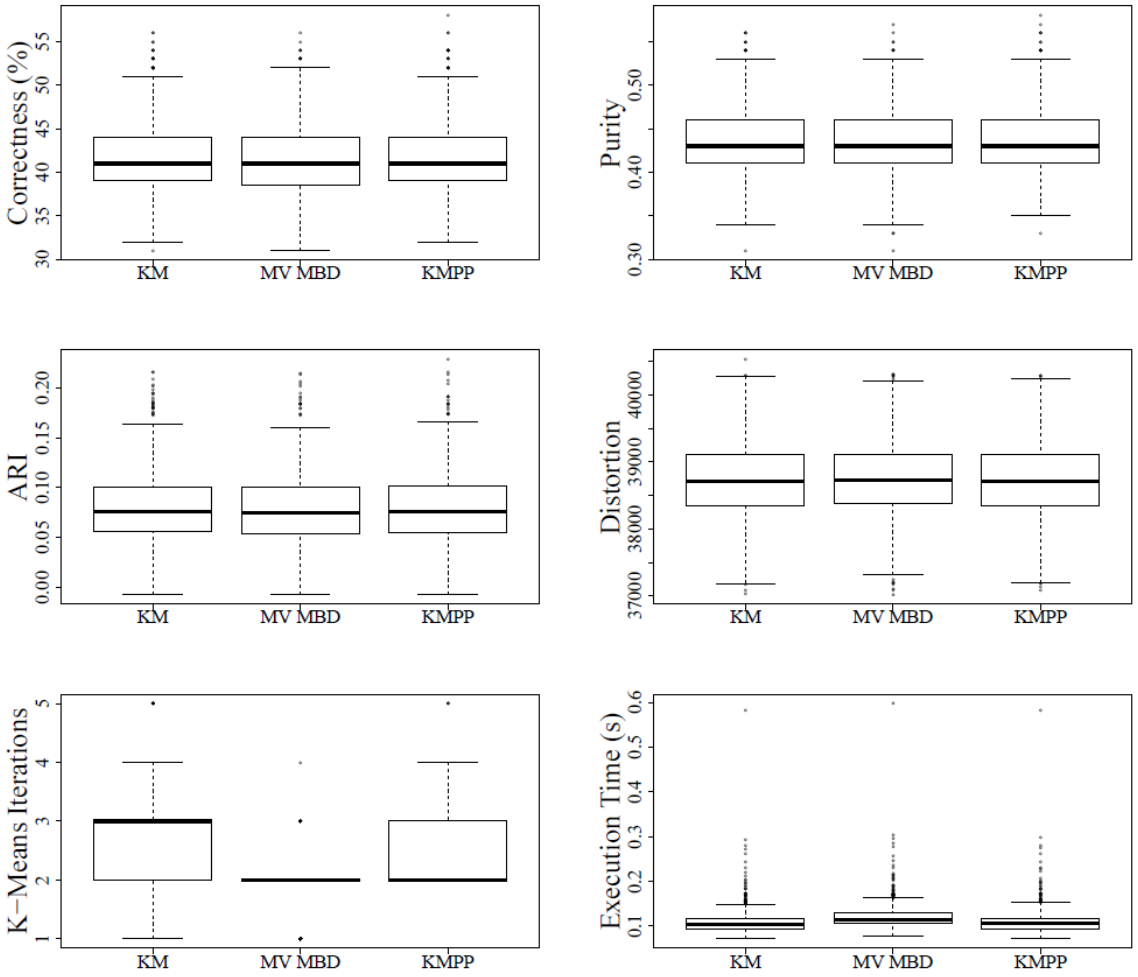


Fig. A.7. Model 1, 3-way CoPADIT measures distribution for sigma = 2.

TABLE A.12.
SUMMARY STATISTICS FOR MODEL 1, 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

`Pmiss` = 0.25, `sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.64	0.65	0.3825	8427	4	0.217
	Mean	0.638	0.6516	0.381	8429	3.653	0.2248
	Variance	0.004535	0.003137	0.0062	22480	0.5131	0.0009733
MV MBD	Median	0.65	0.66	0.3973	8430	3	0.2974
	Mean	0.6434	0.6564	0.3948	8432	3.336	0.3088
	Variance	0.003532	0.002422	0.005411	22460	0.6878	0.001452
FMBD	Median	0.78	0.78	0.5537	8513	2	0.1896
	Mean	0.769	0.7732	0.5516	8508	2.001	0.1967
	Variance	0.004472	0.003392	0.005669	22880	0.1051	0.0009345
KMPP	Median	0.64	0.65	0.3817	8426	4	0.22
	Mean	0.6333	0.648	0.3782	8429	3.651	0.2288
	Variance	0.004288	0.002807	0.005614	22220	0.5477	0.001066
FKMPP	Median	0.74	0.74	0.5182	8518	3	0.1108
	Mean	0.726	0.7457	0.5219	8516	2.776	0.1164
	Variance	0.007462	0.003933	0.006164	23250	0.3782	0.0004766

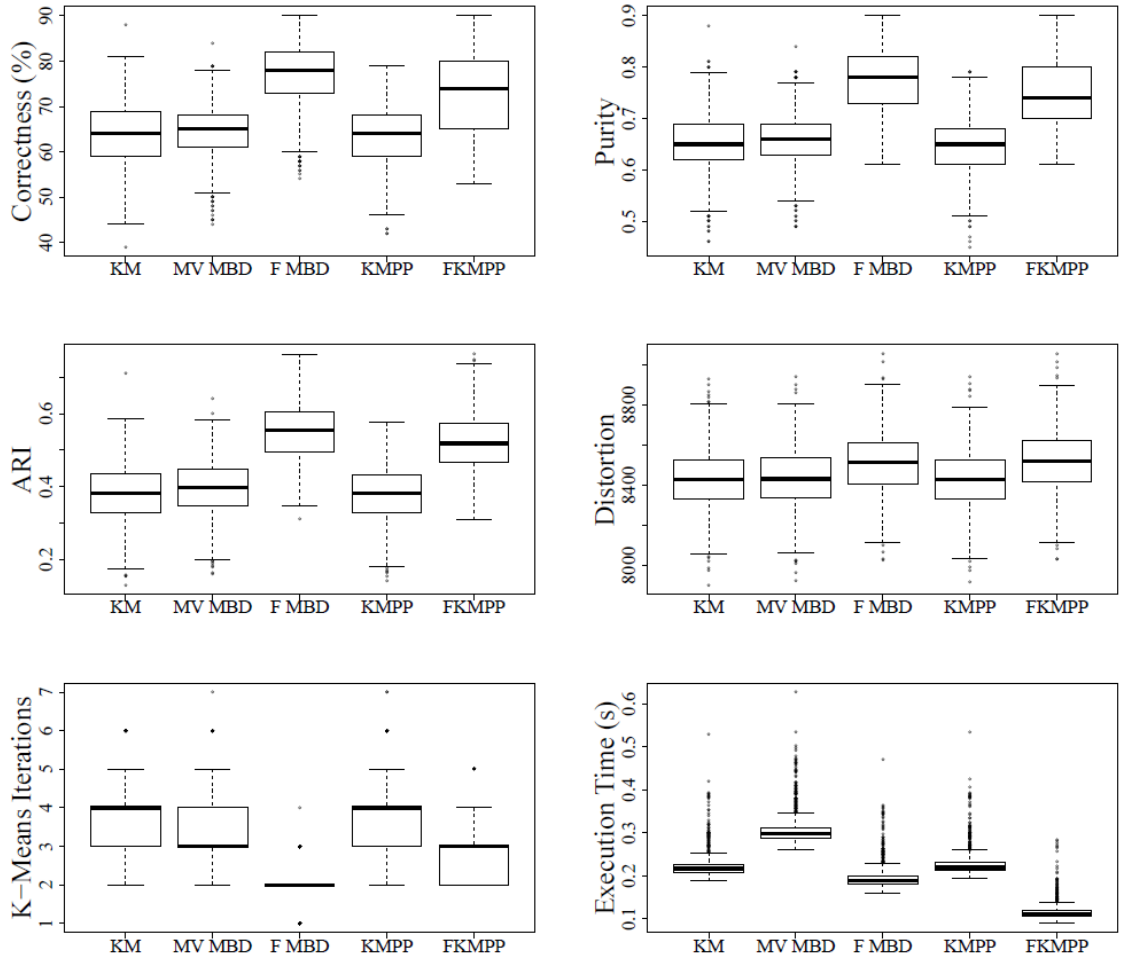


Fig. A.8. Model 1, 25% of missing values, 5-way CoPADIT measures distribution for `sigma` = 1.

TABLE A.13.
SUMMARY STATISTICS FOR MODEL 1, 50% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.5, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.58	0.59	0.2799	7363	4	0.288
	Mean	0.5795	0.5938	0.2819	7358	3.656	0.2988
	Variance	0.004331	0.00309	0.005178	28140	0.5262	0.00156
MV MBD	Median	0.59	0.6	0.2878	7359	3	0.3648
	Mean	0.5853	0.5989	0.2912	7356	3.111	0.3793
	Variance	0.003489	0.002546	0.004678	27580	0.6874	0.002072
FMBD	Median	0.69	0.69	0.4129	7486	2	0.1848
	Mean	0.6866	0.6928	0.4167	7478	2.095	0.1921
	Variance	0.004241	0.003201	0.005269	29010	0.1321	0.0006867
KMPP	Median	0.58	0.59	0.279	7362	4	0.2898
	Mean	0.5786	0.5914	0.2794	7357	3.675	0.3012
	Variance	0.003868	0.002917	0.004788	27190	0.5519	0.001585
FKMPP	Median	0.66	0.67	0.3961	7488	3	0.1047
	Mean	0.6605	0.6767	0.3992	7485	3.035	0.1101
	Variance	0.005522	0.003287	0.005434	28760	0.3641	0.000379

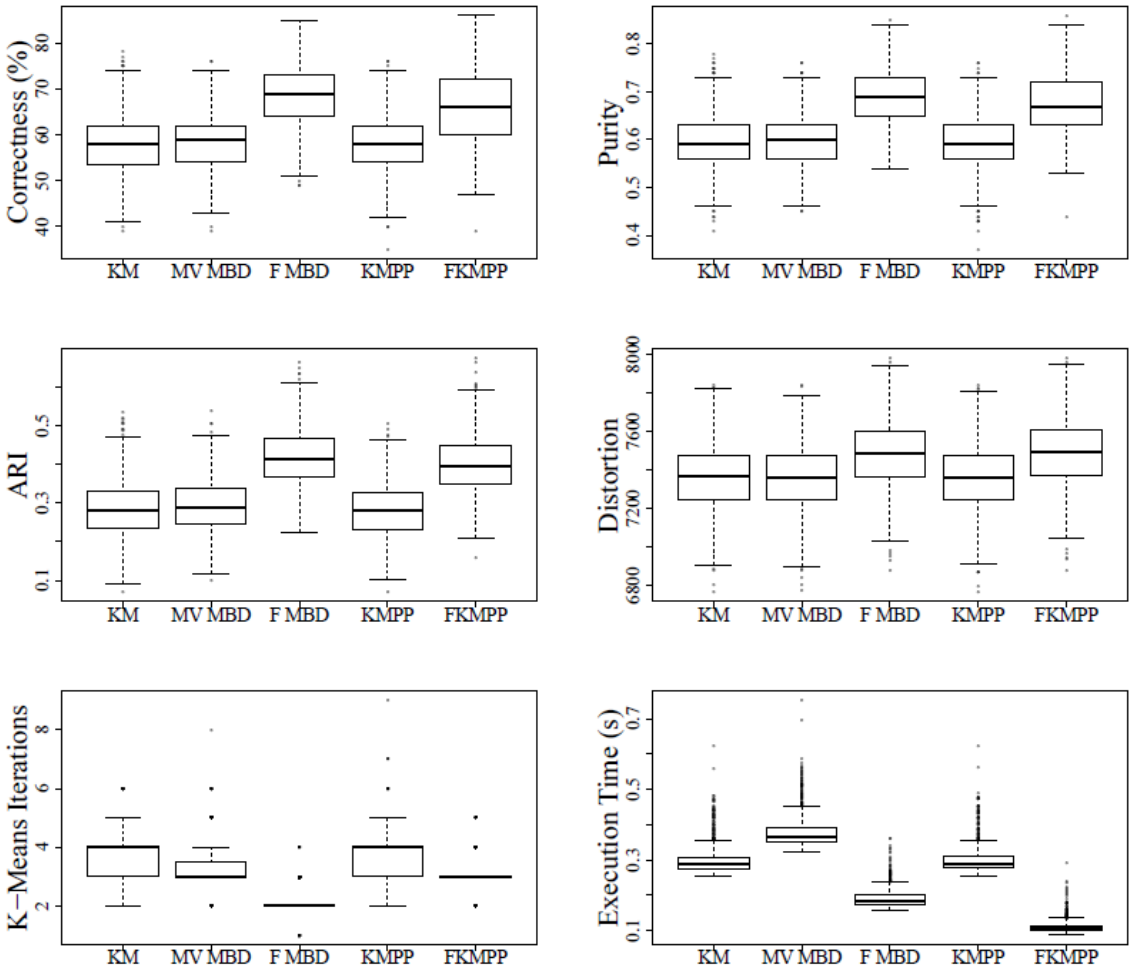


Fig. A.9. Model 1, 50% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.14.
SUMMARY STATISTICS FOR MODEL 1, 75% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.75, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.48	0.5	0.1447	6241	3	0.2169
	Mean	0.4843	0.499	0.1487	6244	3.496	0.2259
	Variance	0.00306	0.002443	0.002807	42540	0.4444	0.001325
MV MBD	Median	0.49	0.51	0.1506	6226	3	0.2931
	Mean	0.4906	0.5052	0.1552	6231	2.698	0.3059
	Variance	0.003117	0.002523	0.002973	42410	0.5213	0.00214
FMBD	Median	0.56	0.57	0.223	6459	2	0.1807
	Mean	0.5551	0.567	0.2273	6465	2.142	0.1892
	Variance	0.003678	0.00273	0.003746	48780	0.194	0.0009213
KMPP	Median	0.48	0.5	0.1458	6232	3	0.2188
	Mean	0.4858	0.5001	0.1488	6240	3.545	0.2285
	Variance	0.003269	0.00269	0.003015	41890	0.4965	0.001321
FKMPP	Median	0.54	0.56	0.221	6473	3	0.1055
	Mean	0.546	0.5608	0.2224	6478	3.15	0.1107
	Variance	0.00393	0.002797	0.003791	50020	0.3959	0.0004074

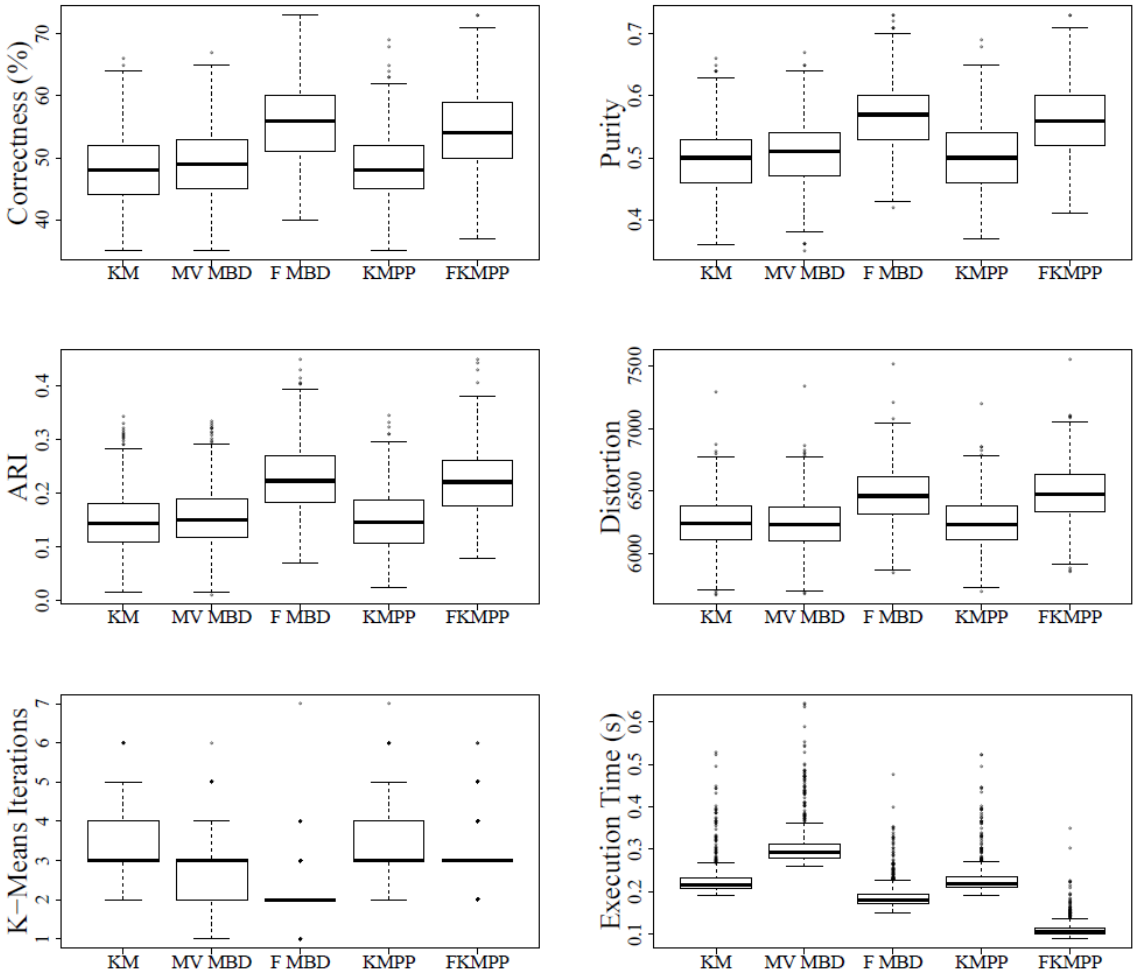


Fig. A.10. Model 1, 75% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.15.
SUMMARY STATISTICS FOR MODEL 2, 5-WAY COPADIT FOR SIGMA = 0.

`sigma` = 0							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	1	1	1	0	1	~ 0
	Mean	1	1	1	0	1	9.146e-04
	Variance	0	0	0	0	0	3.673e-06
MV MBD	Median	1	1	1	0	1	0.07809
	Mean	1	1	1	0	1	0.08003
	Variance	0	0	0	0	0	0.0001103
FMBD	Median	1	1	1	0	1	0.2119
	Mean	1	1	1	0	1	0.2188
	Variance	0	0	0	0	0	0.0005277
KMPP	Median	1	1	1	0	1	2.656e-03
	Mean	1	1	1	0	1	3.382e-03
	Variance	0	0	0	0	0	2.508e-05
FKMPP	Median	1	1	1	0	1	0.137
	Mean	1	1	1	0	1	0.1417
	Variance	0	0	0	0	0	0.0003177

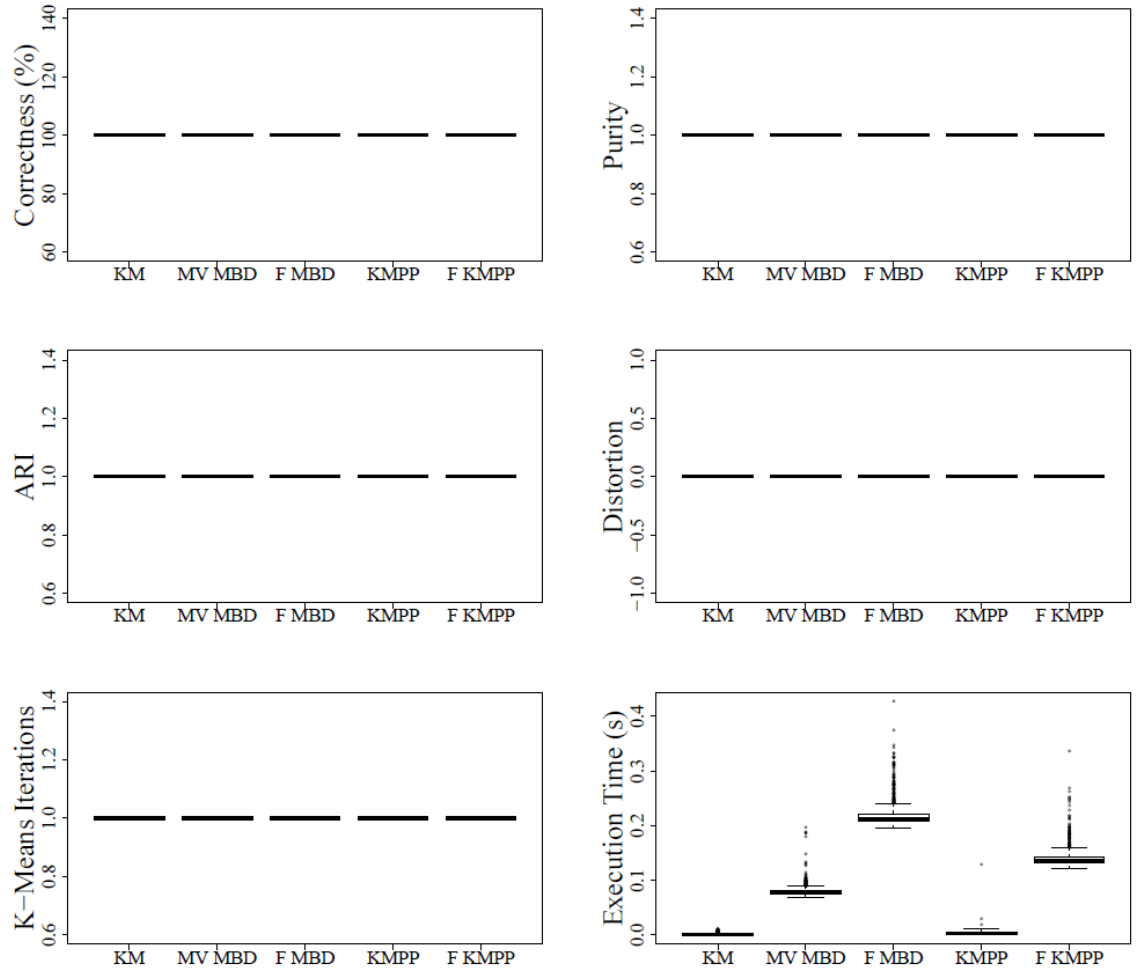


Fig. A.11. Model 2, 5-way CoPADIT measures distribution for `sigma` = 0.

TABLE A.16.
SUMMARY STATISTICS FOR MODEL 2, 5-WAY COPADIT FOR SIGMA = 0.5.

`sigma` = 0.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.715	0.75	0.6669	2849	2	~ 0
	Mean	0.815	0.866	0.7993	2758	2	7.593e-04
	Variance	0.03649	0.01956	0.04456	140600	0.1862	3.591e-06
MV MBD	Median	1	1	1	2421	1	0.07331
	Mean	1	1	1	2422	1	0.0777
	Variance	0	0	0	1299	0	0.0003254
FMBD	Median	1	1	1	2421	1	0.21
	Mean	1	1	1	2422	1	0.2214
	Variance	0	0	0	1299	0	0.00121
KMPP	Median	1	1	1	2421	2	0.002993
	Mean	0.8501	0.8912	0.8378	2422	1.8120	0.003648
	Variance	0.03227	0.01713	0.03845	1299	0.2229	0.0000116
FKMPP	Median	1	1	1	2436	1	0.1407
	Mean	0.9245	0.9194	0.9456	2548	1.424	0.1489
	Variance	0.02078	0.02363	0.0176	60960	0.2565	0.0008039

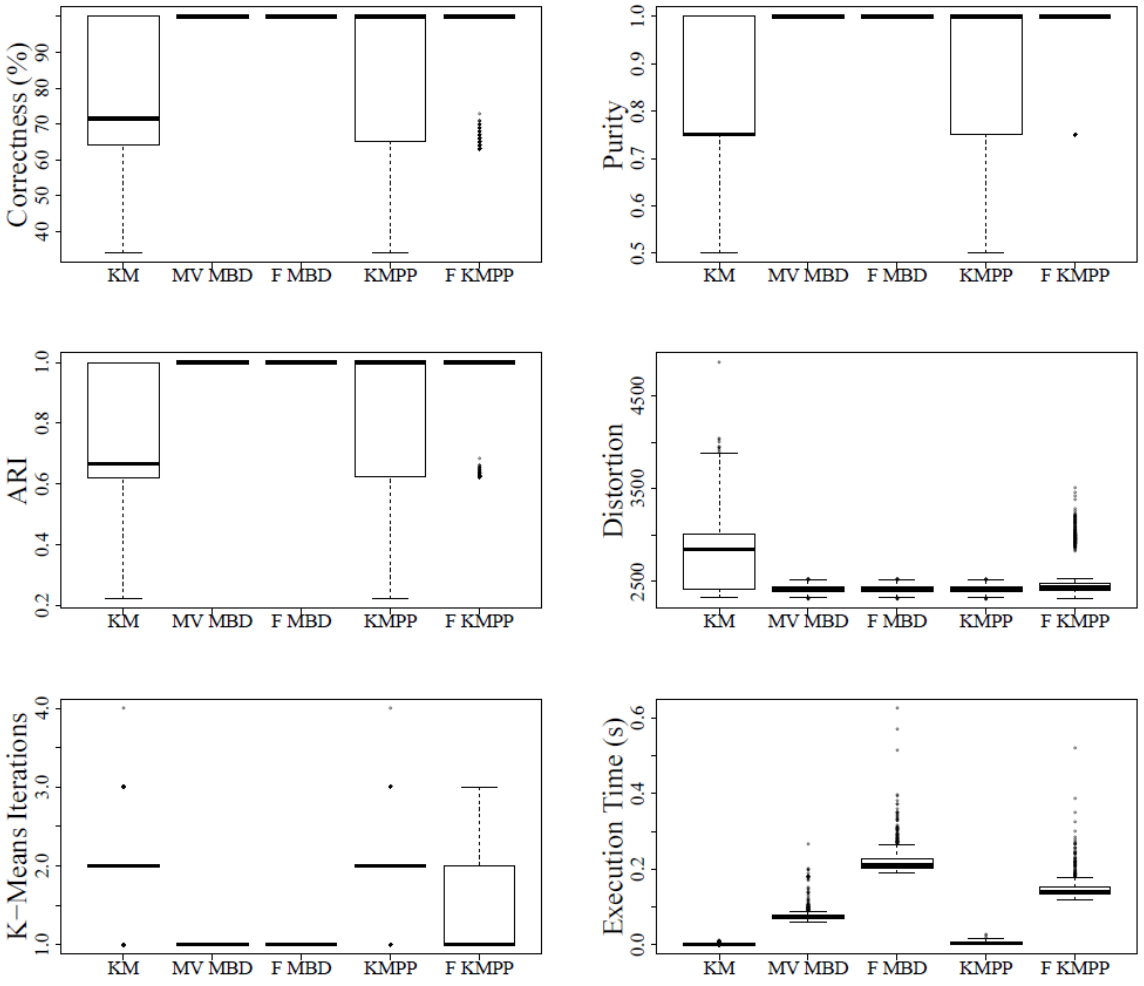


Fig. A.12. Model 2, 5-way CoPADIT measures distribution for `sigma` = 0.5.

TABLE A.17.
SUMMARY STATISTICS FOR MODEL 2, 5-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	1	1	1	9744	2	~ 0
	Mean	0.9236	0.9444	0.9137	9799	2.518	7.927e-04
	Variance	0.02038	0.01062	0.02481	70330	0.31	3.412e-06
MV MBD	Median	1	1	1	9683	1	0.07834
	Mean	0.9983	0.9984	0.9961	9687	1.301	0.08138
	Variance	0.0001429	7.61e-05	0.000243	21060	0.2166	0.0002188
FMBD	Median	1	1	1	9684	1	0.2251
	Mean	0.9994	0.994	0.9984	9687	1.023	0.2378
	Variance	5.934e-06	5.93e-06	4.3e-05	20780	0.02249	0.001664
KMPP	Median	1	1	1	9755	2	3.036e-03
	Mean	0.91	0.9341	0.8986	9821	2.455	3.783e-03
	Variance	0.02321	0.01228	0.02825	76280	0.2943	1.219e-05
FKMPP	Median	1	1	1	9735	2	0.1502
	Mean	0.934	0.9522	0.927	9791	2.09	0.1611
	Variance	0.01847	0.009576	0.02195	68310	0.2722	0.001262

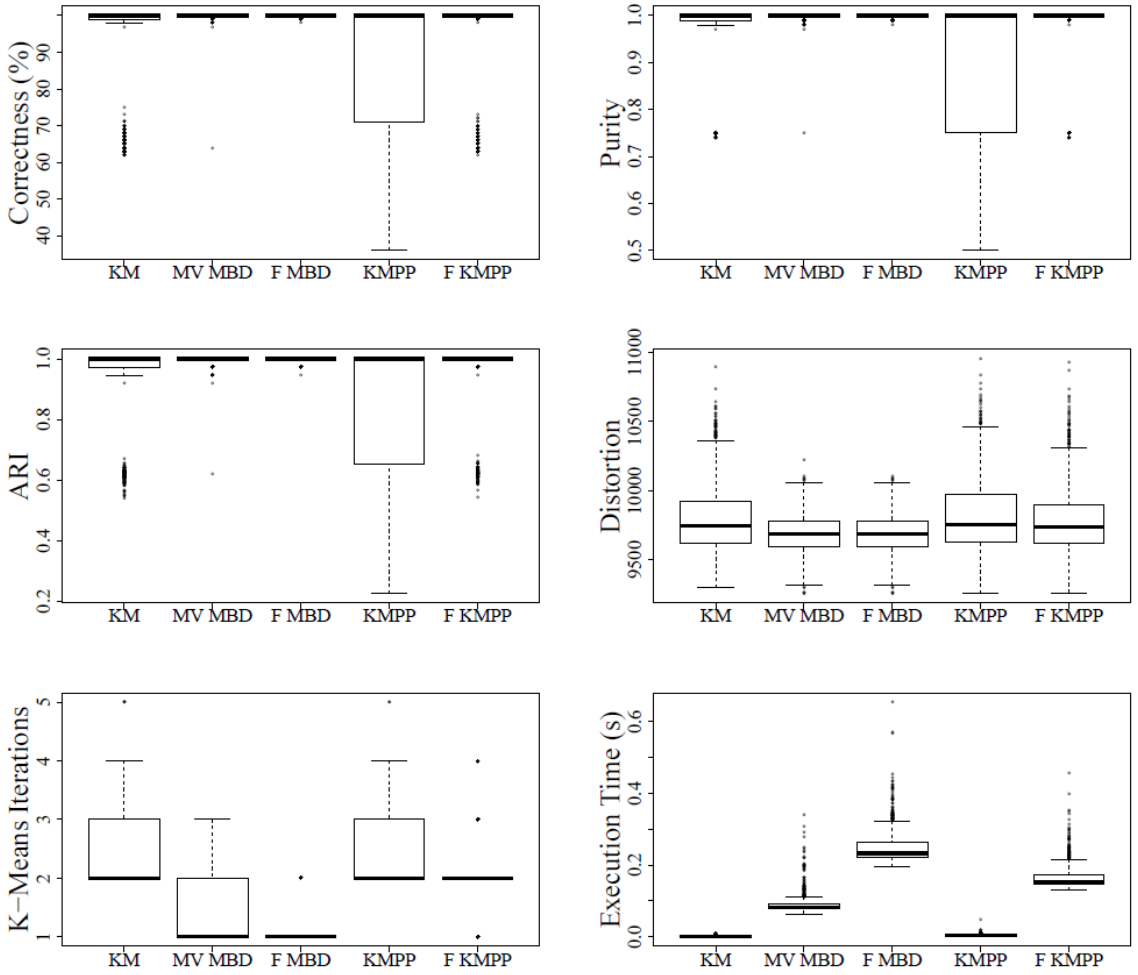


Fig. A.13. Model 2, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.18.
SUMMARY STATISTICS FOR MODEL 2, 5-WAY COPADIT FOR SIGMA = 1.5.

sigma = 1.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.95	0.95	0.8731	21790	3	~ 0
	Mean	0.9182	0.9254	0.8387	21800	3.285	8.086e-04
	Variance	0.009332	0.00605	0.01639	116500	0.3561	3.313e-06
MV MBD	Median	0.96	0.96	0.894	21780	3	0.07978
	Mean	0.9332	0.9365	0.8583	21780	2.629	0.08338
	Variance	0.005845	0.004295	0.01132	112300	0.5199	0.0003244
FMBD	Median	0.98	0.98	0.9467	21780	1	0.2274
	Mean	0.9794	0.9794	0.9456	21780	1.4890	0.2364
	Variance	0.0002173	0.000217	0.00145	105500	0.2501	0.001187
KMPP	Median	0.95	0.95	0.8732	21800	3	3.564e-03
	Mean	0.9168	0.9249	0.8388	21800	3.297	4.066e-03
	Variance	0.009704	0.006086	0.01589	114700	0.3712	1.211e-05
FKMPP	Median	0.98	0.98	0.9467	21820	3	0.1494
	Mean	0.9468	0.9556	0.9087	21830	2.648	0.1564
	Variance	0.009687	0.005718	0.01325	124500	0.3364	0.0007069

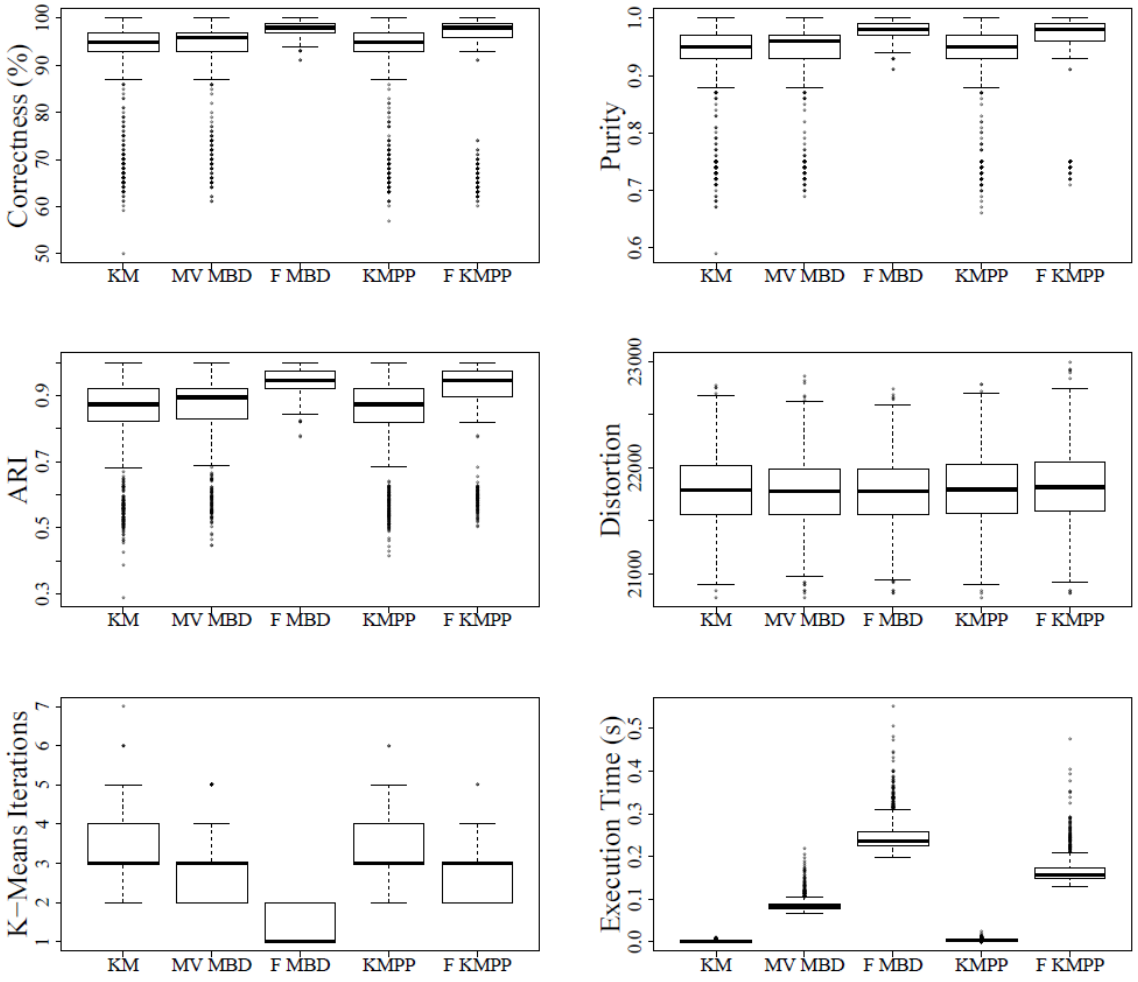


Fig. A.14. Model 2, 5-way CoPADIT measures distribution for sigma = 1.5.

TABLE A.19.
SUMMARY STATISTICS FOR MODEL 2, 5-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.73	0.73	0.4701	38530	4	~ 0
	Mean	0.7225	0.7311	0.4808	38540	3.703	9.378e-04
	Variance	0.009113	0.007412	0.01428	336000	0.5253	3.552e-06
MV MBD	Median	0.72	0.72	0.4763	38560	3	0.08077
	Mean	0.7221	0.7322	0.486	38550	3.4670	0.08269
	Variance	0.008511	0.006688	0.01312	337500	0.6736	0.0001671
FMBD	Median	0.92	0.92	0.7897	38650	2	0.2242
	Mean	0.913	0.913	0.7847	38650	1.983	0.2309
	Variance	0.001077	0.001051	0.005202	331000	0.09881	0.0007395
KMPP	Median	0.72	0.72	0.467	38520	4	3.131e-03
	Mean	0.7189	0.7287	0.4778	38540	3.756	3.787e-03
	Variance	0.009167	0.007188	0.01384	332900	0.6171	1.145e-05
FKMPP	Median	0.91	0.91	0.7731	38680	3	0.1473
	Mean	0.8844	0.8907	0.7514	38680	3.145	0.1523
	Variance	0.007374	0.004756	0.01282	344700	0.3643	0.0005088

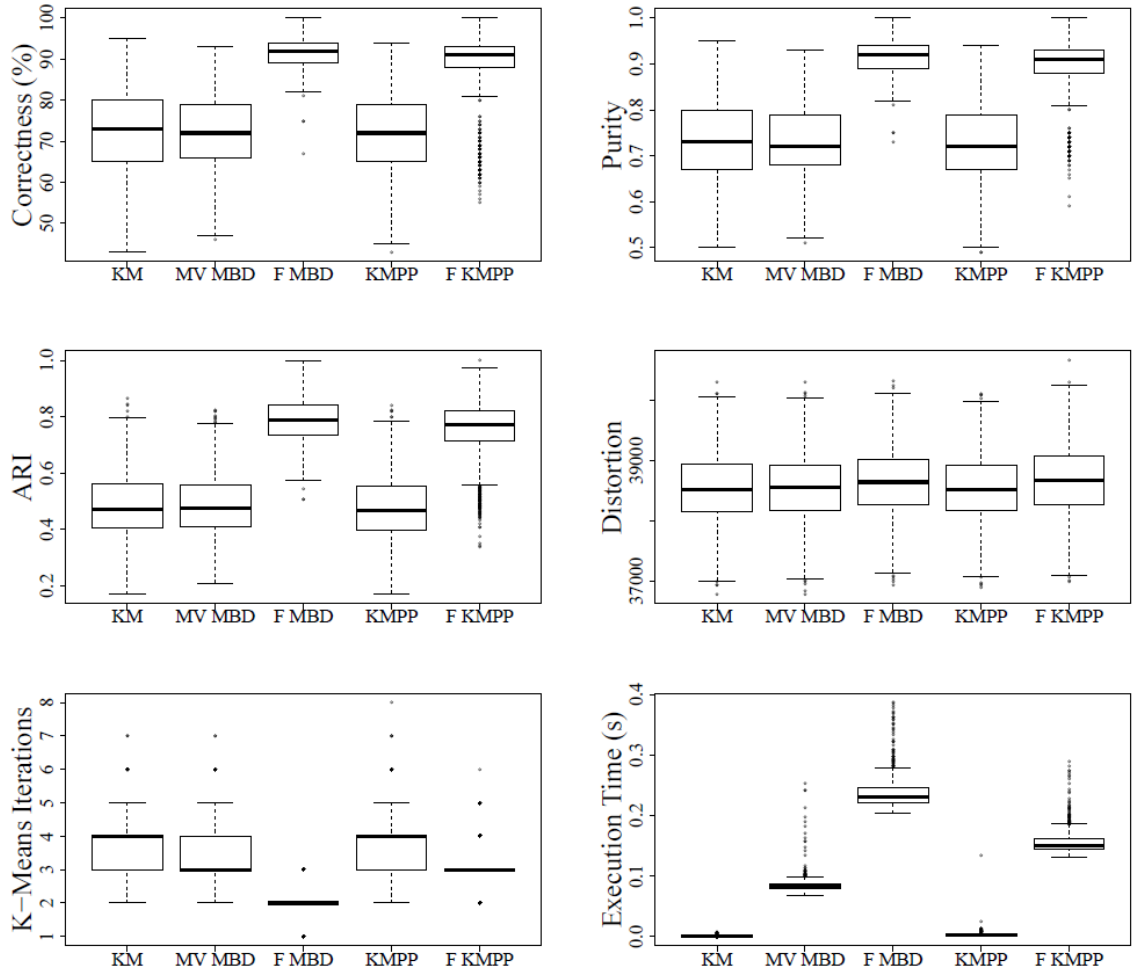


Fig. A.15. Model 2, 5-way CoPADIT measures distribution for sigma = 2.

The p-values for the paired t-test of correctness, purity and ARI for all methods are collected in the following tables for the different values of `sigma`.

TABLE A.20.
MODEL 2 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 0.5.

`sigma` = 0.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	9.03e-146	9.03e-146	3.26e-05	3.30e-43
	Purity		1.563e-143	1.563e-143	4.376e-05	1.303e-43
	ARI		5.433e-142	5.433e-142	3.404e-05	1.658e-43
MV MBD	Correctness	9.03e-146	-	-	7.746e-117	1.385e-54
	Purity	1.563e-143		-	4.133e-116	8.793e-55
	ARI	5.433e-142		-	2.466e-115	9.551e-55
FMBD	Correctness	9.03e-146	-	-	7.746e-117	1.385e-54
	Purity	1.563e-143	-		4.133e-116	8.793e-55
	ARI	5.433e-142	-		2.466e-115	9.551e-55
KMPP	Correctness	3.26e-05	7.746e-117	7.746e-117	-	4.760e-24
	Purity	4.376e-05	4.133e-116	4.133e-116		8.115e-25
	ARI	3.404e-05	2.466e-115	2.466e-115		1.268e-24
FKMPP	Correctness	3.30e-43	1.385e-54	1.385e-54	4.760e-24	-
	Purity	1.303e-43	8.793e-55	8.793e-55	8.115e-25	
	ARI	1.658e-43	9.551e-55	9.551e-55	1.268e-24	

TABLE A.21.
MODEL 2 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

`sigma` = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	3.151e-54	5.144e-56	3.869e-02	1.033e-01
	Purity		1.566e-54	1.302e-56	3.159e-02	8.992e-02
	ARI		2.046e-54	2.282e-57	3.773e-02	5.752e-02
MV MBD	Correctness	3.151e-54	-	4.426e-03	5.017e-65	2.134e-45
	Purity	1.566e-54		4.265e-04	4.087e-65	2.869e-45
	ARI	2.046e-54		1.707e-06	5.775e-65	4.246e-44
FMBD	Correctness	5.144e-56	4.426e-03	-	2.472e-66	3.481e-47
	Purity	1.302e-56	4.265e-04		9.908e-67	2.104e-47
	ARI	2.282e-57	1.707e-06		1.459e-67	2.577e-47
KMPP	Correctness	3.869e-02	5.017e-65	2.472e-66	-	1.374e-04
	Purity	3.159e-02	4.087e-65	9.908e-67		7.364e-05
	ARI	3.773e-02	5.775e-65	1.459e-67		4.232e-05
FKMPP	Correctness	1.033e-01	2.134e-45	3.481e-47	1.374e-04	-
	Purity	8.992e-02	2.869e-45	2.104e-47	7.364e-05	
	ARI	5.752e-02	4.246e-44	2.577e-47	4.232e-05	

TABLE A.22.
MODEL 2 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.5.

sigma = 1.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	3.440e-05	1.237e-75	7.491e-01	4.428e-11
	Purity		1.666e-04	3.071e-88	8.711e-01	1.830e-19
	ARI		1.577e-05	7.672e-119	9.868e-01	1.589e-37
MV MBD	Correctness	3.440e-05	-	5.157e-71	9.473e-06	3.194e-04
	Purity	1.666e-04		1.107e-81	8.350e-05	9.324e-11
	ARI	1.577e-05		2.492e-119	1.358e-05	1.080e-26
FMBD	Correctness	1.237e-75	5.157e-71	-	1.818e-76	9.826e-25
	Purity	3.071e-88	1.107e-81		6.891e-90	5.968e-25
	ARI	7.672e-119	2.492e-119		2.256e-123	6.606e-25
KMPP	Correctness	7.491e-01	9.473e-06	1.818e-76		8.880e-13
	Purity	8.711e-01	8.350e-05	6.891e-90		1.248e-21
	ARI	9.868e-01	1.358e-05	2.256e-123		5.231e-41
FKMPP	Correctness	4.428e-11	3.194e-04	9.826e-25	8.880e-13	-
	Purity	1.830e-19	9.324e-11	5.968e-25	1.248e-21	
	ARI	1.589e-37	1.080e-26	6.606e-25	5.231e-41	

TABLE A.23.
MODEL 2 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 2.

sigma = 2						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	8.976e-01	~ 0	4.540e-01	4.327e-219
	Purity		7.294e-01	~ 0	4.540e-01	4.475e-262
	ARI		2.173e-01	~ 0	4.584e-01	5.459e-311
MV MBD	Correctness	8.976e-01	-	~ 0	3.646e-01	2.536e-220
	Purity	7.294e-01		~ 0	2.553e-01	3.919e-265
	ARI	2.173e-01		~ 0	4.048e-02	7.659e-309
FMBD	Correctness	~ 0	~ 0	-	~ 0	2.619e-26
	Purity	~ 0	~ 0		~ 0	1.436e-26
	ARI	~ 0	~ 0		~ 0	1.686e-25
KMPP	Correctness	3.179e-01	3.646e-01	~ 0	-	3.600e-224
	Purity	4.540e-01	2.553e-01	~ 0		2.959e-269
	ARI	4.584e-01	4.048e-02	~ 0		7.152e-317
FKMPP	Correctness	4.327e-219	2.536e-220	2.619e-26	3.600e-224	-
	Purity	4.475e-262	3.919e-265	1.436e-26	2.959e-269	
	ARI	5.459e-311	7.659e-309	1.686e-25	7.152e-317	

TABLE A.24.
SUMMARY STATISTICS FOR MODEL 2, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 1.

`sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.77	0.77	0.6268	10290	3	0.1312
	Mean	0.7741	0.7859	0.6299	10300	3.155	0.1411
	Variance	0.006251	0.004161	0.008532	63700	0.4574	0.001368
MV MBD	Median	0.78	0.78	0.6352	10260	2	0.1521
	Mean	0.7912	0.7959	0.6432	10260	2.139	0.16
	Variance	0.004404	0.003566	0.006472	50580	0.1779	0.00157
KMPP	Median	0.77	0.77	0.6272	10290	3	0.1368
	Mean	0.775	0.7868	0.6316	10290	3.116	0.1429
	Variance	0.00569	0.003654	0.007323	57720	0.477	0.001401

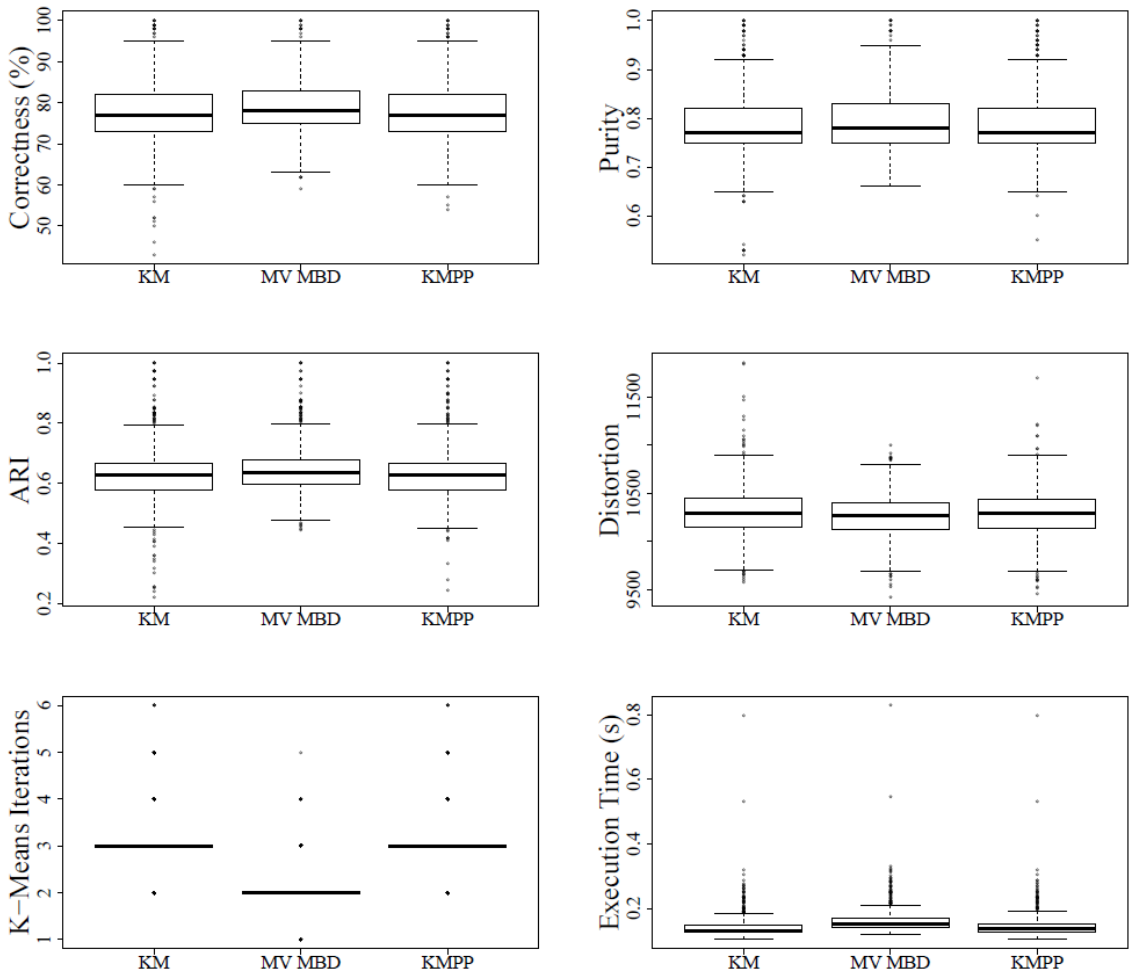


Fig. A.16. Model 2, 3-way CoPADIT measures distribution for `sigma` = 1.

TABLE A.25.
SUMMARY STATISTICS FOR MODEL 2, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.38	0.39	0.03714	40770	3	0.1346
	Mean	0.3876	0.3979	0.04647	40760	3.345	0.1394
	Variance	0.002305	0.002201	0.001716	398000	0.4304	0.0006278
MV MBD	Median	0.38	0.395	0.04054	40740	2	0.1532
	Mean	0.3881	0.3988	0.0469	40770	2.29	0.1579
	Variance	0.002153	0.002089	0.001594	401000	0.2321	0.0007203
KMPP	Median	0.38	0.4	0.04172	40750	3	0.137
	Mean	0.3907	0.4017	0.04943	40750	3.285	0.1411
	Variance	0.002207	0.002105	0.001658	395000	0.3681	0.0006635

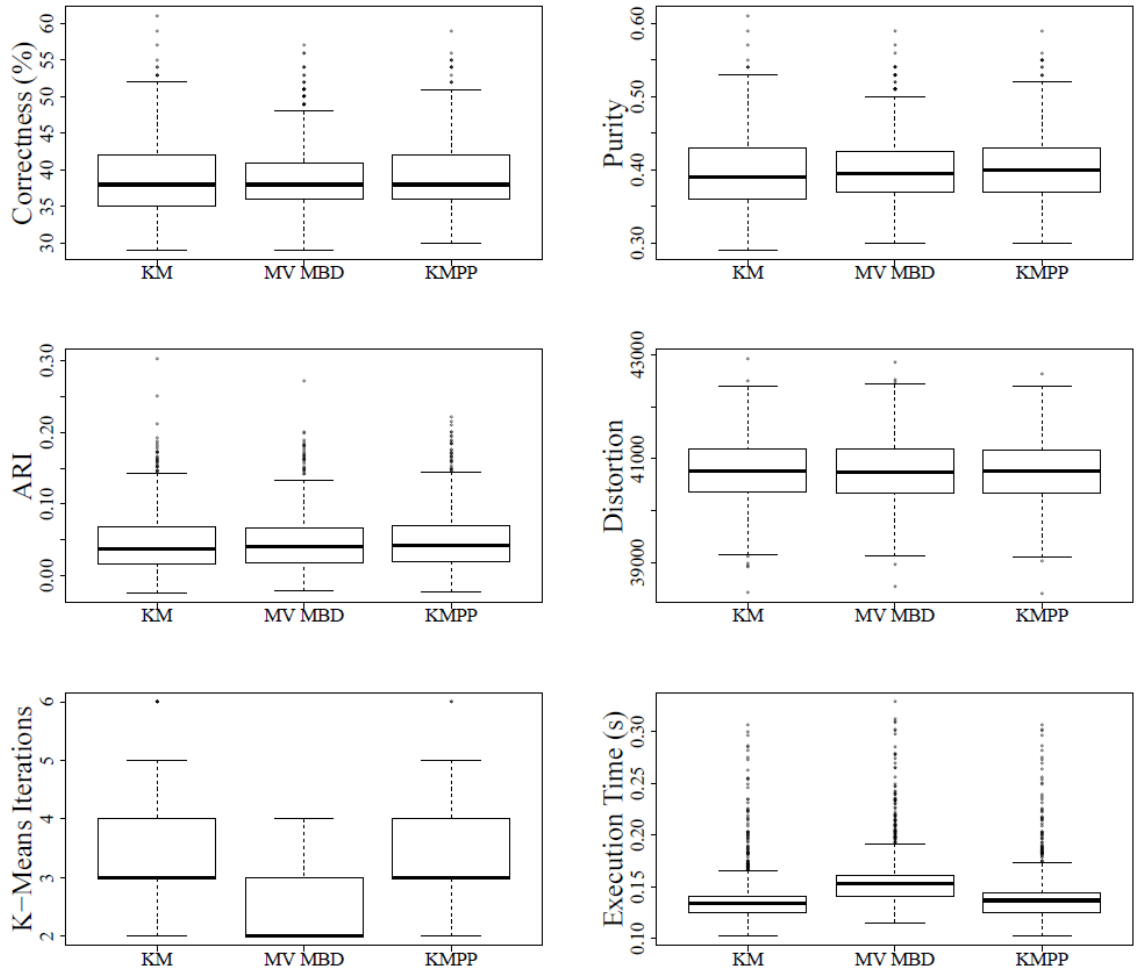


Fig. A.17. Model 2, 3-way CoPADIT measures distribution for sigma = 2.

TABLE A.26.
SUMMARY STATISTICS FOR MODEL 2, 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

`Pmiss` = 0.25, `sigma` = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.99	0.99	0.9731	8607	3	0.1952
	Mean	0.933	0.9493	0.9152	8660	2.635	0.2094
	Variance	0.01682	0.008974	0.0213	66530	0.3701	0.004659
MV MBD	Median	1	1	1	8572	1	0.265
	Mean	0.9938	0.9938	0.9835	8571	1.469	0.2835
	Variance	6,23E-02	6,23E-02	0.000442	23140	0.2593	0.008379
FMBD	Median	1	1	1	8574	1	0.2036
	Mean	0.9961	0.9961	0.9896	8572	1.131	0.2158
	Variance	4,04E-02	4,04E-02	0.000288	23150	0.114	0.005168
KMPP	Median	0.99	0.99	0.9731	8608	3	0.197
	Mean	0.9383	0.9531	0.9216	8650	2.564	0.2119
	Variance	0.01573	0.008439	0.01983	59670	0.3102	0.00479
FKMPP	Median	1	1	1	8606	2	0.134
	Mean	0.9419	0.9568	0.9298	8655	2.319	0.1437
	Variance	0.01584	0.008329	0.01949	63940	0.2695	0.002646

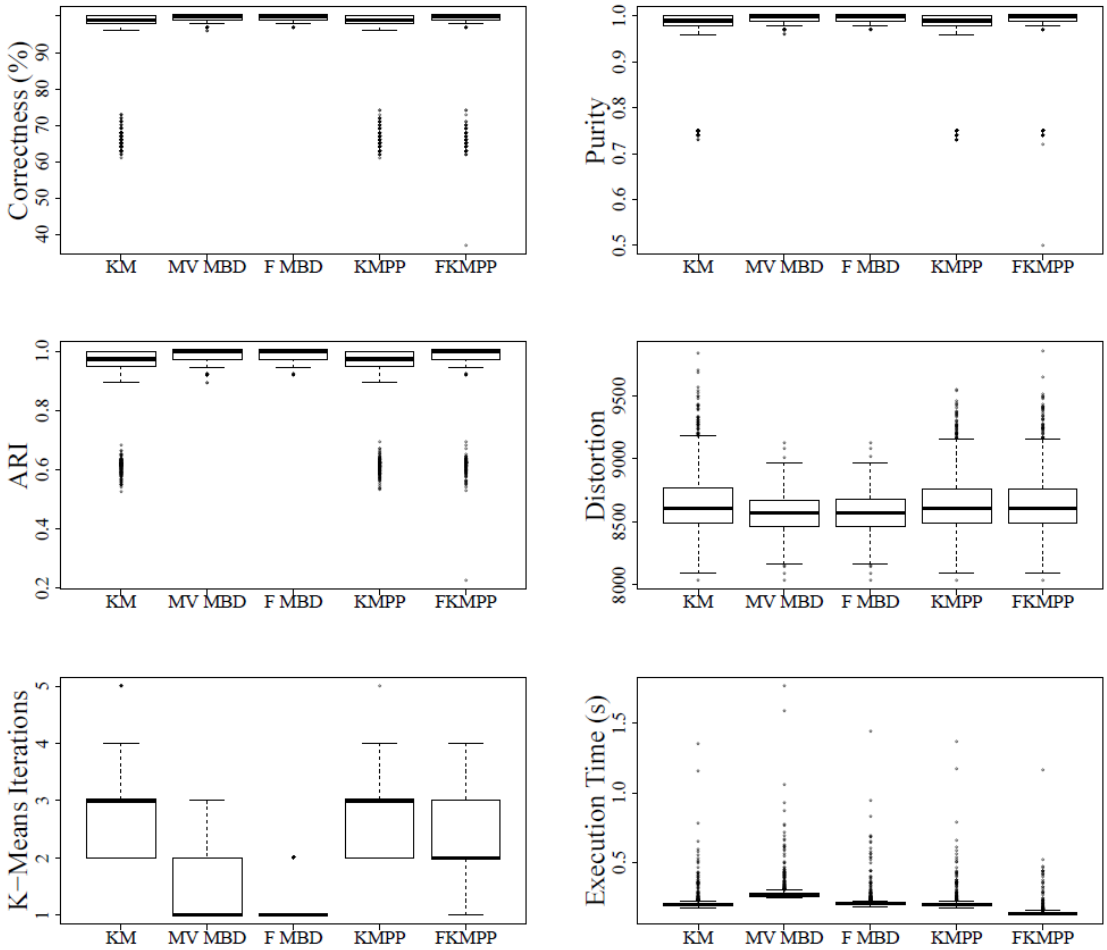


Fig. A.18. Model 2, 25% of missing values, 5-way CoPADIT measures distribution for `sigma` = 1.

TABLE A.27.
SUMMARY STATISTICS FOR MODEL 2, 50% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.5, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.97	0.97	0.9203	7625	3	0.2593
	Mean	0.9364	0.9452	0.8837	7641	2.857	0.2644
	Variance	0.009746	0.005389	0.01378	48420	0.3349	0.0005006
MV MBD	Median	0.97	0.97	0.9214	7600	2	0.3325
	Mean	0.9688	0.9689	0.9199	7598	1.838	0.3377
	Variance	0.0005865	0.000506	0.002294	29540	0.1539	0.0005728
FMBD	Median	0.96	0.96	0.895	7723	2	0.1935
	Mean	0.8904	0.8924	0.8219	7788	1.667	0.1937
	Variance	0.01509	0.0144	0.02485	105800	0.2464	9,06E-02
KMPP	Median	0.97	0.97	0.9203	7626	3	0.2637
	Mean	0.9341	0.9437	0.8819	7643	2.824	0.267
	Variance	0.01041	0.005716	0.01443	46690	0.3233	0.0005226
FKMPP	Median	0.96	0.96	0.894	7728	3	0.1258
	Mean	0.8872	0.893	0.8158	7792	2.706	0.1284
	Variance	0.01551	0.01332	0.02538	105100	0.388	8.47e-05

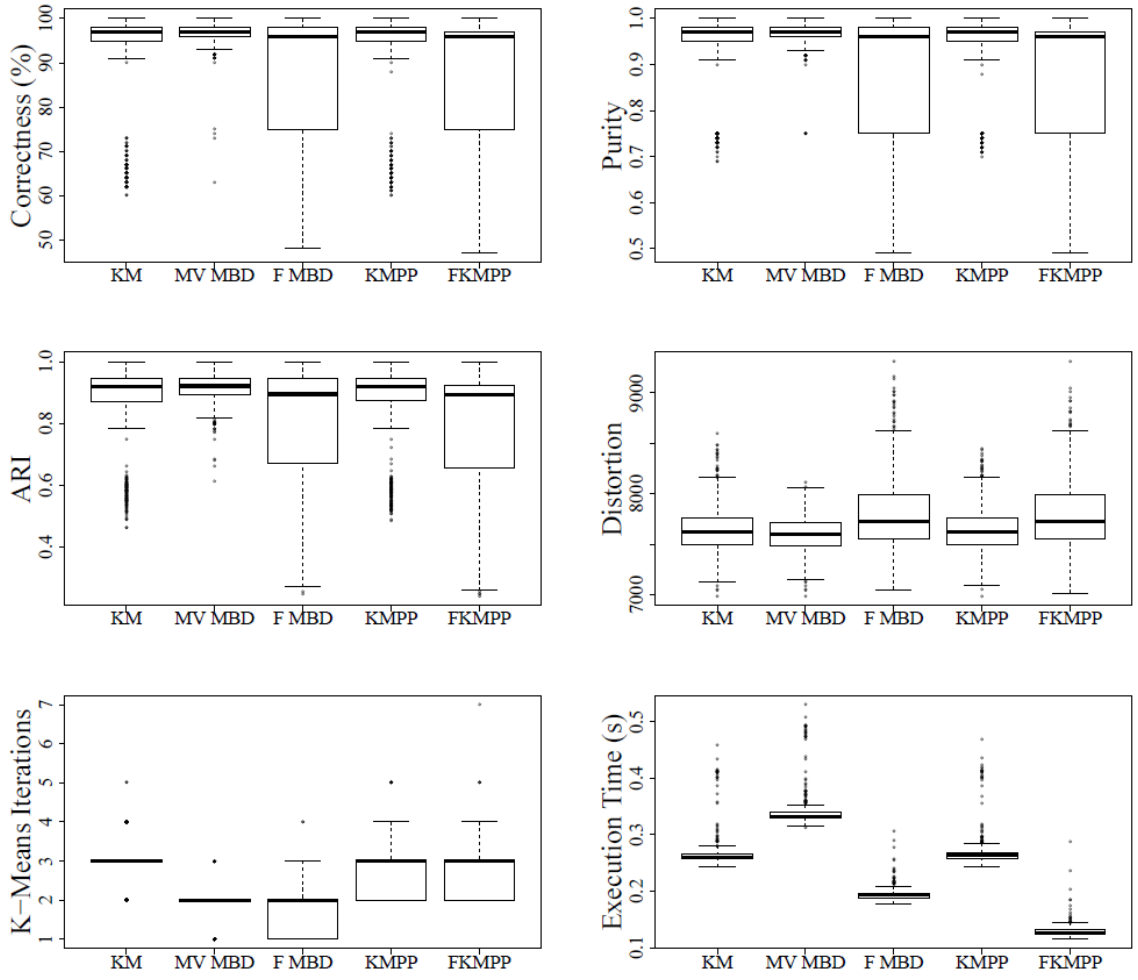


Fig. A.19. Model 2, 50% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.28.
SUMMARY STATISTICS FOR MODEL 2, 75% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.75, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.84	0.84	0.6311	6768	3	0.2291
	Mean	0.818	0.8237	0.6177	6776	3.237	0.2436
	Variance	0.007553	0.005582	0.01278	55420	0.4613	0.002281
MV MBD	Median	0.85	0.85	0.6462	6752	2	0.3095
	Mean	0.841	0.8426	0.6423	6757	2.165	0.3264
	Variance	0.00373	0.003142	0.008826	51590	0.204	0.003222
FMBD	Median	0.27	0.28	0.000152	9606	1	0.2336
	Mean	0.2739	0.2802	9,02E-03	9617	1.085	0.2494
	Variance	3,86E-02	5,54E-03	2,24E-04	110600	0.07787	0.002918
KMPP	Median	0.84	0.84	0.6311	6774	3	0.2314
	Mean	0.8134	0.8204	0.6123	6781	3.252	0.2466
	Variance	0.008092	0.005669	0.01313	55530	0.463	0.002355
FKMPP	Median	0.27	0.28	0.000152	9606	1	0.1547
	Mean	0.2739	0.2802	1.1e-05	9616	1.299	0.1672
	Variance	3,81E-02	5,37E-03	2,25E-04	110200	0.2178	0.00172

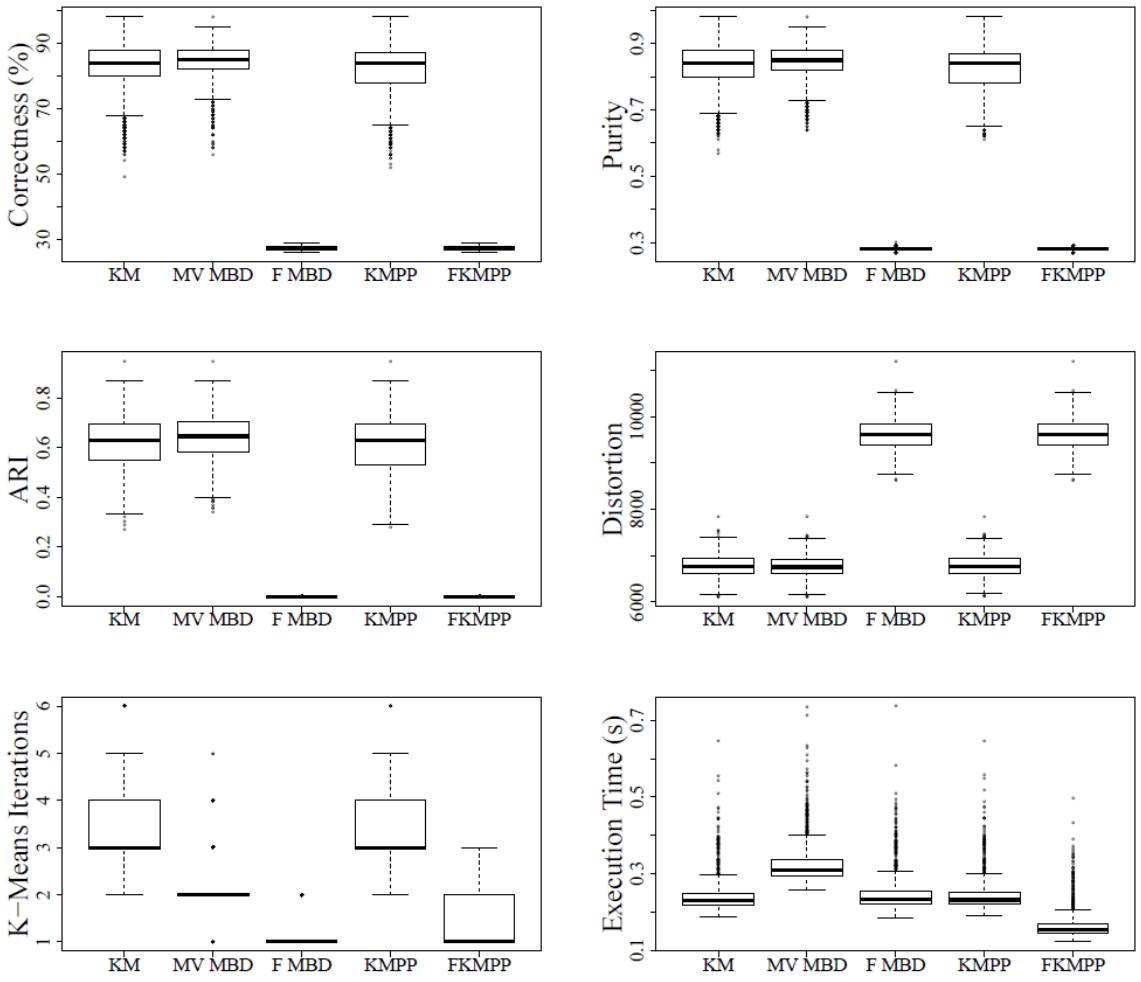


Fig. A.20. Model 2, 75% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

Model Three - Five-Way Comparison

TABLE A.29.
SUMMARY STATISTICS FOR MODEL 3, 5-WAY COPADIT FOR SIGMA = 0.

sigma = 0							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	1	1	1	0	1	~ 0
	Mean	1	1	1	0	1	1.988e-03
	Variance	0	0	0	0	0	4.109e-05
MV MBD	Median	1	1	1	0	1	0.1944
	Mean	1	1	1	0	1	0.2023
	Variance	0	0	0	0	0	0.001151
FMBD	Median	1	1	1	0	1	0.3941
	Mean	1	1	1	0	1	0.4072
	Variance	0	0	0	0	0	0.003958
KMPP	Median	1	1	1	0	1	1.018e-02
	Mean	1	1	1	0	1	9.727e-03
	Variance	0	0	0	0	0	2.121e-05
FKMPP	Median	1	1	1	0	1	0.2026
	Mean	1	1	1	0	1	0.2133
	Variance	0	0	0	0	0	0.001731

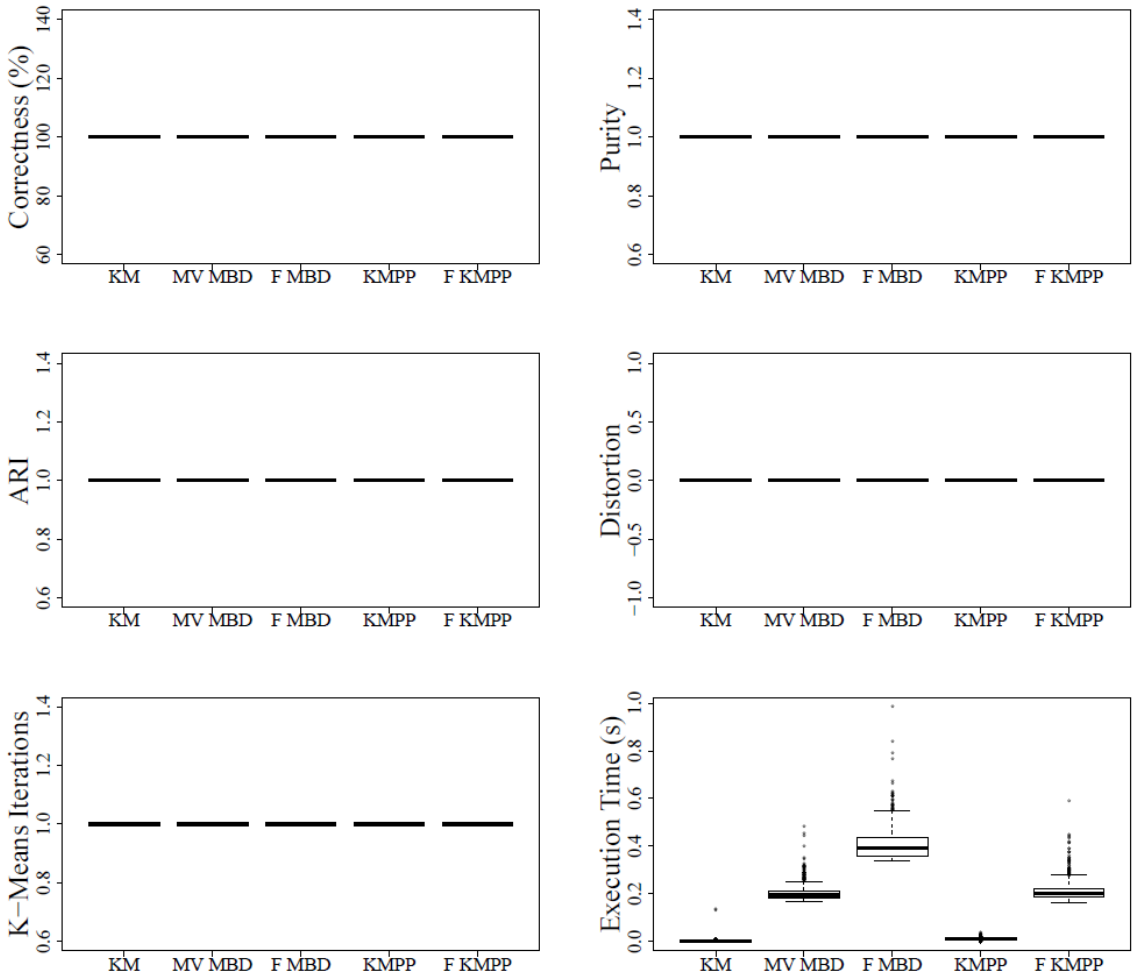


Fig. A.21. Model 3, 5-way CoPADIT measures distribution for sigma = 0.

TABLE A.30.
SUMMARY STATISTICS FOR MODEL 3, 5-WAY COPADIT FOR SIGMA = 0.5.

`sigma` = 0.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.568	0.6	0.4051	5979	3	~ 0
	Mean	0.5601	0.5996	0.3977	5980	3.508	1.685e-03
	Variance	0.005812	0.003236	0.003649	3404	0.5825	8.781e-06
MV MBD	Median	0.568	0.592	0.4162	5982	4	0.186
	Mean	0.5611	0.5946	0.4089	5981	3.685	0.1913
	Variance	0.00452	0.002734	0.00304	3446	0.6704	0.0004094
FMBD	Median	0.832	0.832	0.6541	6017	2	0.3484
	Mean	0.8085	0.8206	0.6495	6015	2.089	0.3551
	Variance	0.007274	0.004016	0.006346	3353	0.1012	0.0006999
KMPP	Median	0.56	0.6	0.4036	5980	3	1.014e-02
	Mean	0.5586	0.5961	0.3949	5980	3.493	9.480e-03
	Variance	0.005399	0.003092	0.00348	3367	0.5825	3.731e-05
FKMPP	Median	0.688	0.736	0.5691	6022	3	0.1834
	Mean	0.7112	0.7509	0.5813	6023	2.821	0.1867
	Variance	0.01453	0.007684	0.009322	3537	0.3733	0.0003799

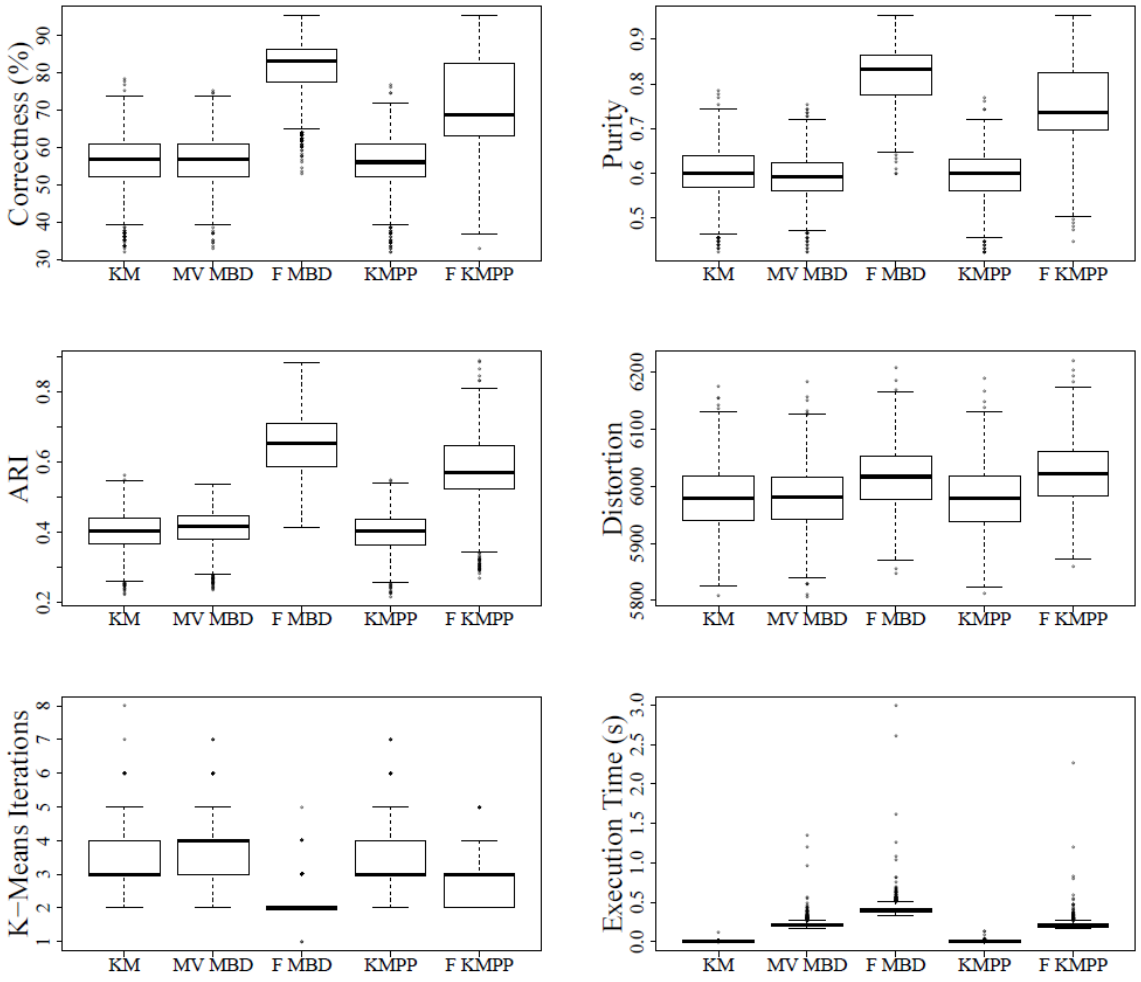


Fig. A.22. Model 3, 5-way CoPADIT measures distribution for `sigma` = 0.5.

TABLE A.31.
SUMMARY STATISTICS FOR MODEL 3, 5-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.44	0.448	0.2151	23680	4	1.994e-03
	Mean	0.4408	0.455	0.2176	23680	4.116	2.388e-03
	Variance	0.001321	0.001027	0.001423	52170	0.7273	8.160e-06
MV MBD	Median	0.432	0.452	0.2424	23680	4	0.2207
	Mean	0.4361	0.4535	0.243	23690	4.169	0.2312
	Variance	0.001049	0.000899	0.001571	50760	0.7672	0.001464
FMBD	Median	0.536	0.552	0.3336	23970	3	0.4112
	Mean	0.5382	0.5591	0.3336	23970	2.59	0.4244
	Variance	0.003123	0.002053	0.002033	52540	0.428	0.002837
KMPP	Median	0.44	0.448	0.2123	23680	4	0.01022
	Mean	0.4396	0.4533	0.215	23680	4.113	0.01238
	Variance	0.001263	0.000979	0.001298	51590	0.739	0.0001091
FKMPP	Median	0.52	0.552	0.3166	23980	3	0.2137
	Mean	0.5236	0.552	0.3175	23980	3.393	0.224
	Variance	0.003387	0.002134	0.002406	53700	0.465	0.001465

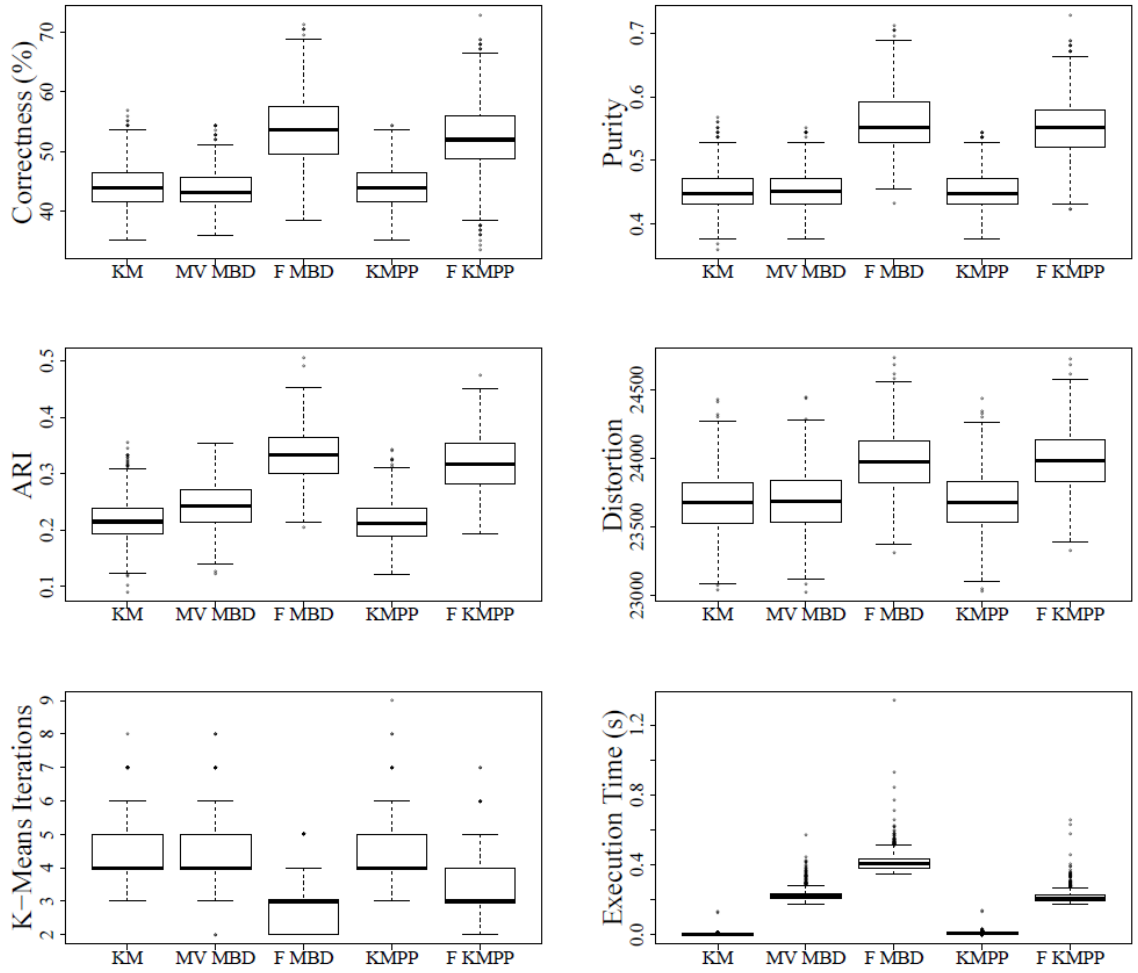


Fig. A.23. Model 3, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.32.
SUMMARY STATISTICS FOR MODEL 3, 5-WAY COPADIT FOR SIGMA = 1.5.

sigma = 1.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.352	0.368	0.07075	52990	4	1.995e-03
	Mean	0.3562	0.3676	0.07365	52970	4.46	2.438e-03
	Variance	0.001154	0.001148	0.001391	254900	0.8973	8.523e-06
MV MBD	Median	0.36	0.368	0.08227	52980	5	0.2374
	Mean	0.3593	0.371	0.08324	52970	4.684	0.2531
	Variance	0.0009899	0.001043	0.001593	257000	0.9651	0.003616
FMBD	Median	0.448	0.464	0.1965	53790	3	0.4122
	Mean	0.4474	0.4619	0.1983	53800	2.818	0.4433
	Variance	0.001645	0.001376	0.001408	263400	0.4854	0.01292
KMPP	Median	0.352	0.368	0.06702	52980	4	0.01039
	Mean	0.3531	0.3648	0.0711	52960	4.512	0.01274
	Variance	0.001257	0.001179	0.001335	256500	0.8968	0.0001435
FKMPP	Median	0.44	0.456	0.1898	53800	4	0.2133
	Mean	0.4434	0.4593	0.192	53810	3.721	0.2338
	Variance	0.001752	0.00142	0.001385	267900	0.6098	0.006608

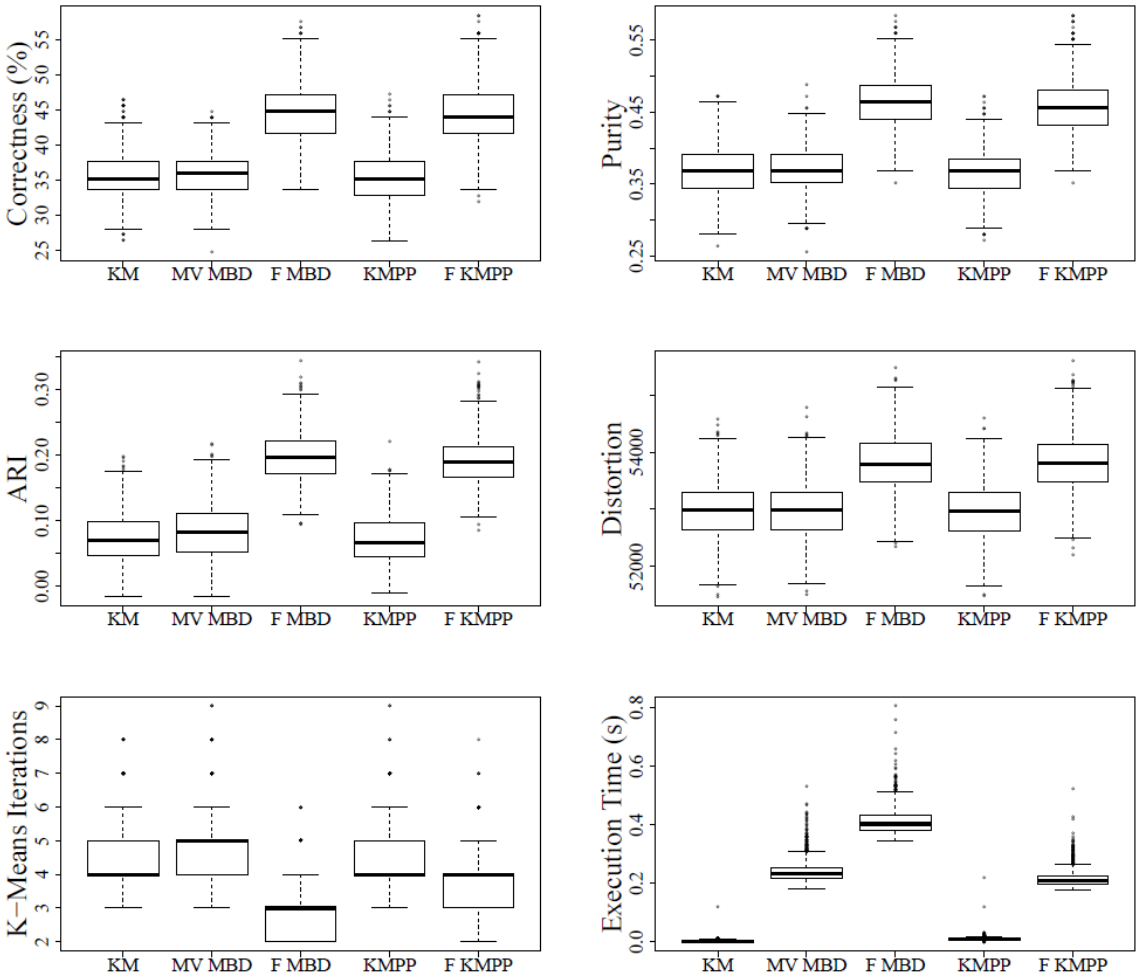


Fig. A.24. Model 3, 5-way CoPADIT measures distribution for sigma = 1.5

TABLE A.33.
SUMMARY STATISTICS FOR MODEL 3, 5-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.312	0.328	0.02074	93820	4	~ 0
	Mean	0.3156	0.3253	0.02405	93810	4.436	2.476e-03
	Variance	0.0007859	0.000798	0.000485	804100	0.8588	3.089e-05
MV MBD	Median	0.312	0.328	0.02433	93840	5	0.216
	Mean	0.3169	0.3273	0.02758	93820	4.764	0.2229
	Variance	0.0008079	0.000842	0.000556	797600	0.9673	0.001484
FMBD	Median	0.392	0.408	0.119	95500	3	0.3704
	Mean	0.3948	0.4068	0.1206	95500	2.912	0.3863
	Variance	0.001226	0.00108	0.00106	831600	0.4667	0.00289
KMPP	Median	0.312	0.328	0.02139	93820	4	1.561e-02
	Mean	0.3159	0.3257	0.02438	93800	4.402	9.274e-03
	Variance	0.0007562	0.000785	0.000491	796700	0.8853	7.875e-05
FKMPP	Median	0.392	0.4	0.1154	95530	4	0.1881
	Mean	0.3921	0.4045	0.1165	95520	3.913	0.1999
	Variance	0.001234	0.001134	0.001121	849600	0.6801	0.001115

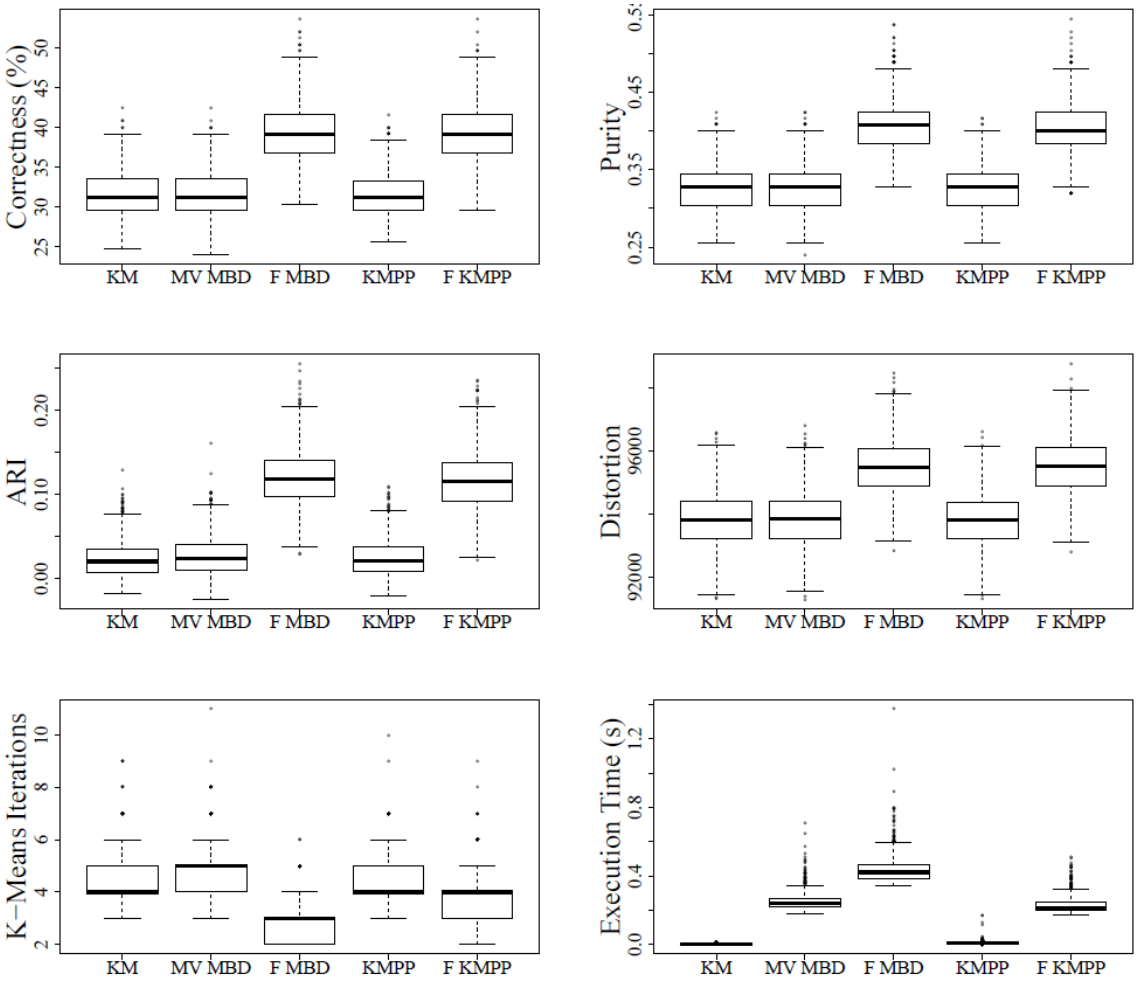


Fig. A.25. Model 3, 5-way CoPADIT measures distribution for sigma = 2

The p-values for the paired t-test of correctness, purity and ARI for all methods are collected in the following tables for the different values of `sigma`.

TABLE A.34.
MODEL 3 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 0.5.

`sigma` = 0.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	6.208e-01	~ 0	6.208e-01	1.216e-169
	Purity		2.534e-02	~ 0	1.286e-01	1.870e-257
	ARI		4.514e-07	~ 0	2.247e-01	2.748e-294
MV MBD	Correctness	7.639e-01	-	~ 0	4.045e-01	1.957e-175
	Purity	2.534e-02		~ 0	5.175e-01	5.759e-273
	ARI	4.514e-07		~ 0	2.774e-10	5.272e-283
FMBD	Correctness	~ 0	~ 0	-	~ 0	1.484e-88
	Purity	~ 0	~ 0		~ 0	5.859e-92
	ARI	~ 0	~ 0		~ 0	2.83e-80
KMPP	Correctness	6.208e-01	4.045e-01	~ 0	-	6.052e-170
	Purity	1.286e-01	5.175e-01	~ 0		6.487e-261
	ARI	2.247e-01	2.774e-10	~ 0		8.519e-298
FKMPP	Correctness	1.216e-169	1.957e-175	1.484e-88	6.052e-170	-
	Purity	1.870e-257	5.759e-273	5.859e-92	6.487e-261	
	ARI	2.748e-294	5.272e-283	2.83e-80	8.519e-298	

TABLE A.35.
MODEL 3 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

`sigma` = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	1.375e-03	1.001e-262	4.345e-01	3.527e-204
	Purity		2.327e-01	~ 0	1.866e-01	9.288e-320
	ARI		3.874e-60	~ 0	5.608e-02	4.272e-297
MV MBD	Correctness	1.375e-03	-	9.213e-284	1.347e-02	1.082e-222
	Purity	2.327e-01		~ 0	0.8801	~ 0
	ARI	3.874e-60		5.042e-285	2.132e-72	1.426e-205
FMBD	Correctness	1.001e-262	9.213e-284	-	1.209e-258	2.920e-11
	Purity	~ 0	~ 0		~ 0	6.434e-06
	ARI	~ 0	5.042e-285		~ 0	3.481e-25
KMPP	Correctness	4.345e-01	1.347e-02	1.209e-258	-	7.438e-208
	Purity	1.866e-01	0.8801	~ 0		1.1818e-320
	ARI	5.608e-02	2.132e-72	~ 0		6.420e-306
FKMPP	Correctness	3.527e-204	1.082e-222	2.920e-11	7.438e-208	-
	Purity	9.288e-320	~ 0	6.434e-06	1.1818e-320	
	ARI	4.272e-297	1.426e-205	3.481e-25	6.420e-306	

TABLE A.36.
MODEL 3 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.5.

sigma = 1.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	1.155e-02	2.451e-313	1.800e-02	8.307e-294
	Purity		0.005602	~ 0	0.032440	~ 0
	ARI		7.717e-13	~ 0	4.902e-02	~ 0
MV MBD	Correctness	1.155e-02	-	3.320e-316	1.463e-06	1.389e-296
	Purity	0.005602		~ 0	1.013e-06	~ 0
	ARI	7.717e-13		~ 0	3.017e-19	~ 0
FMBD	Correctness	2.451e-313	3.320e-316	-	~ 0	9.294e-03
	Purity	~ 0	~ 0		~ 0	0.04576
	ARI	~ 0	~ 0		~ 0	2.735e-07
KMPP	Correctness	1.800e-02	1.463e-06	~ 0		7.673e-312
	Purity	0.032440	1.013e-06	~ 0		~ 0
	ARI	4.902e-02	3.017e-19	~ 0		~ 0
FKMPP	Correctness	8.307e-294	1.389e-296	9.294e-03	7.673e-312	-
	Purity	~ 0	~ 0	0.04576	~ 0	
	ARI	~ 0	~ 0	2.735e-07	~ 0	

TABLE A.37.
MODEL 3 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 2.

sigma = 2						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	2.321e-01	3.953e-323	8.052e-01	8.4760-314
	Purity		0.06004	~ 0	0.71120	~ 0
	ARI		2.375e-05	~ 0	6.872e-01	~ 0
MV MBD	Correctness	2.321e-01	-	3.4081e-320	3.431e-01	4.057e-303
	Purity	0.06004		~ 0	1.382e-01	4.552e-319
	ARI	2.375e-05		~ 0	1.595e-04	~ 0
FMBD	Correctness	3.953e-323	3.4081e-320	-	~ 0	2.16e-02
	Purity	~ 0	~ 0		~ 0	0.02828
	ARI	~ 0	~ 0		~ 0	1.305e-05
KMPP	Correctness	8.052e-01	3.431e-01	~ 0		~ 0
	Purity	0.71120	1.382e-01	~ 0		~ 0
	ARI	6.872e-01	1.595e-04	~ 0		~ 0
FKMPP	Correctness	8.4760-314	4.057e-303	2.16e-02	~ 0	-
	Purity	~ 0	4.552e-319	0.02828	~ 0	
	ARI	~ 0	~ 0	1.305e-05	~ 0	

Model Three - Coefficient Clustering

TABLE A.38.
SUMMARY STATISTICS FOR MODEL 3, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.328	0.336	0.02929	24460	4	0.1718
	Mean	0.3282	0.339	0.03587	24460	3.646	0.1751
	Variance	0.001158	0.001178	0.000899	59580	0.5632	0.0003907
MV MBD	Median	0.32	0.336	0.02884	24460	2	0.1885
	Mean	0.3275	0.3379	0.03469	24460	2.479	0.1957
	Variance	0.001204	0.001189	0.000876	57340	0.3699	0.0005232
KMPP	Median	0.328	0.336	0.03261	24460	3	0.1719
	Mean	0.3294	0.3396	0.03596	24460	3.516	0.1782
	Variance	0.001087	0.001072	0.000813	57880	0.4762	0.0004313

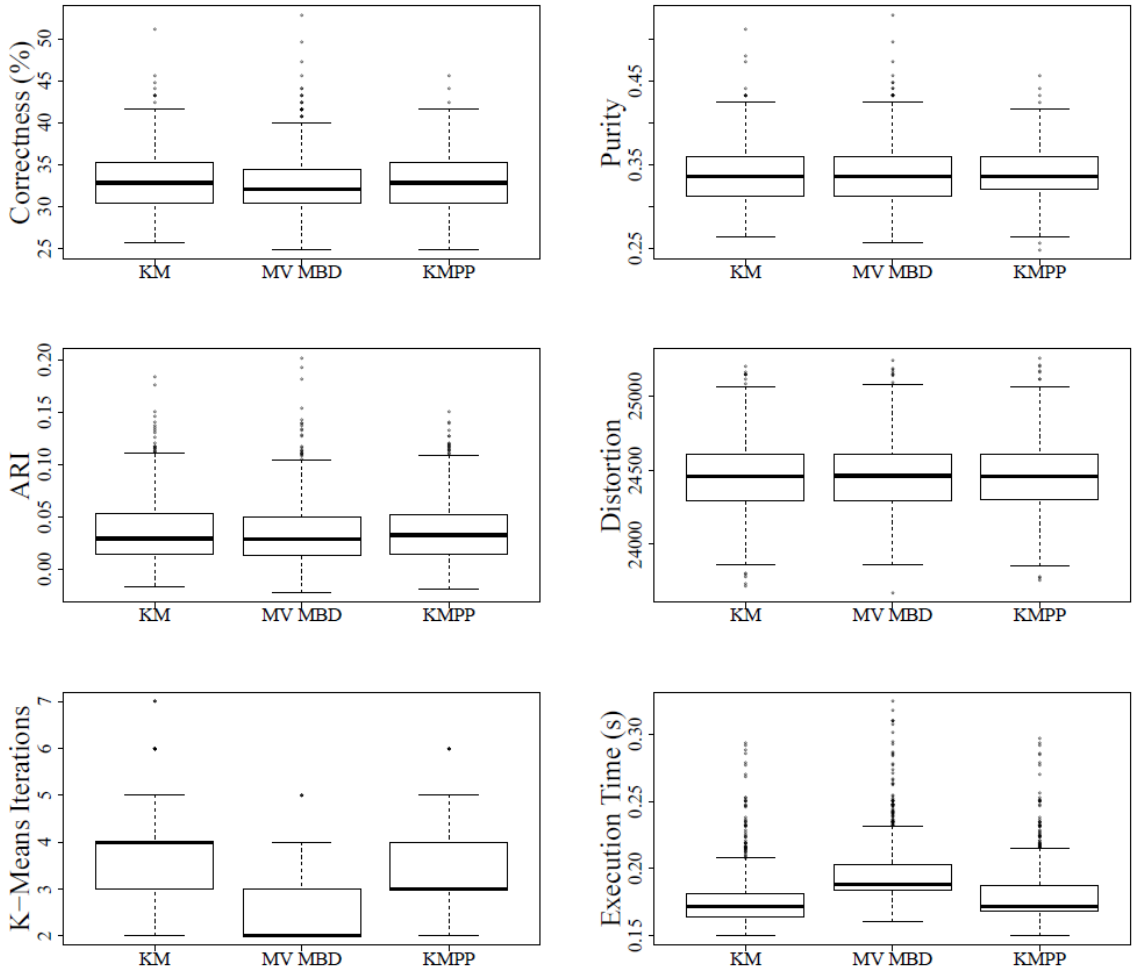


Fig. A.26. Model 3, 3-way CoPADIT measures distribution for sigma = 1.

TABLE A.39.
SUMMARY STATISTICS FOR MODEL 3, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.288	0.296	0.002747	96200	3	0.1718
	Mean	0.2923	0.301	0.004292	96170	3.591	0.1744
	Variance	0.0004923	0.000511	0.000165	842700	0.5723	0.0003405
MV MBD	Median	0.296	0.304	0.003455	96200	2	0.1884
	Mean	0.2937	0.302	0.0046	96180	2.443	0.1948
	Variance	0.0004838	0.000479	0.000156	834300	0.3471	0.0004366
KMPP	Median	0.296	0.304	0.00315	96190	3	0.1719
	Mean	0.2936	0.3021	0.004779	96180	3.511	0.1778
	Variance	0.0004928	0.000504	0.000165	844900	0.5424	0.0004033

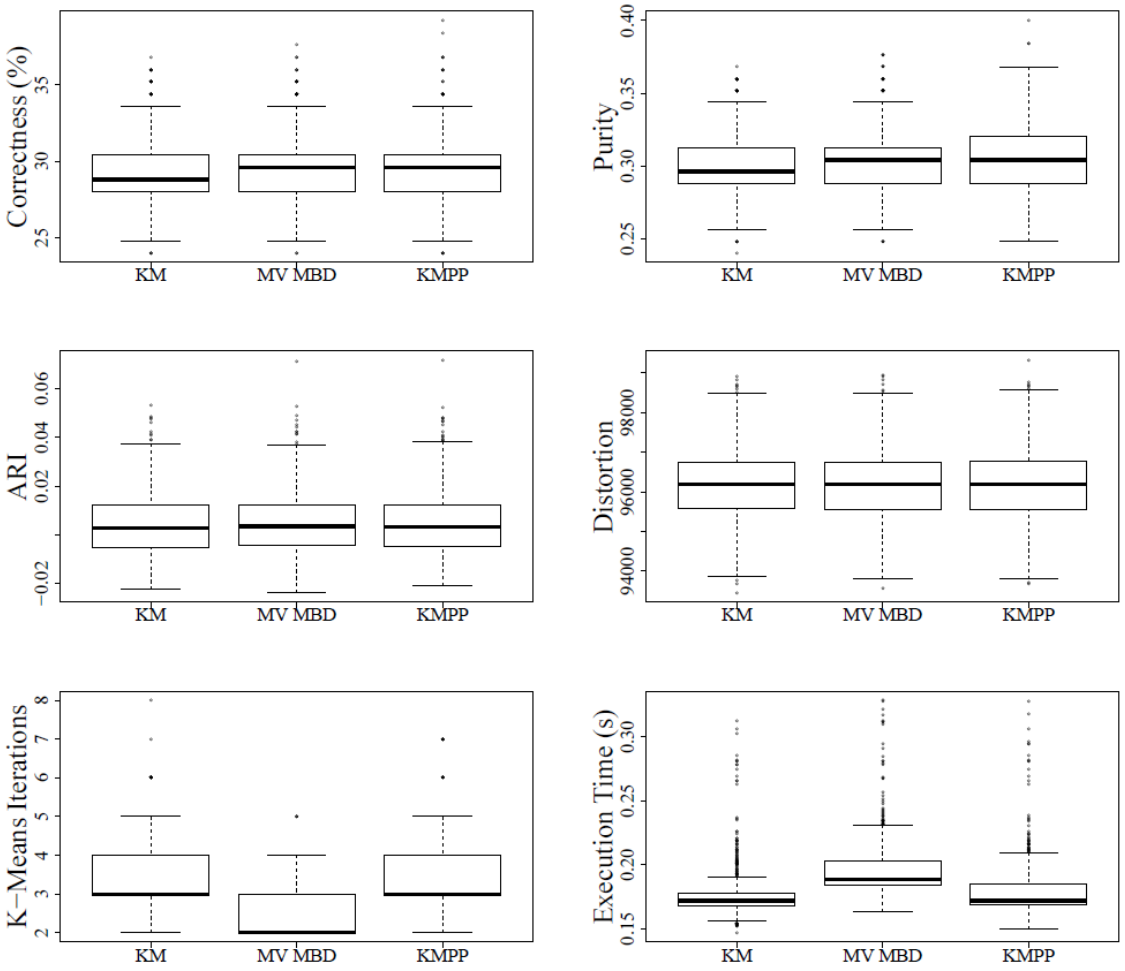


Fig. A.27. Model 3, 3-way CoPADIT measures distribution for sigma = 2.

Model Three - Missing Data

TABLE A.40.
SUMMARY STATISTICS FOR MODEL 3, 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.25, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.432	0.448	0.2012	20850	4	0.5069
	Mean	0.4363	0.4506	0.203	20850	4.091	0.5238
	Variance	0.001355	0.001115	0.001315	55870	0.6514	0.00385
MV MBD	Median	0.432	0.448	0.2221	20860	4	0.7054
	Mean	0.434	0.4503	0.2235	20860	4.108	0.729
	Variance	0.001202	0.000987	0.001589	56310	0.7811	0.006569
FMBD	Median	0.496	0.52	0.2831	21090	3	0.3721
	Mean	0.5011	0.5199	0.2841	21090	2.674	0.3923
	Variance	0.002354	0.001743	0.001873	56690	0.4702	0.003326
KMPP	Median	0.432	0.448	0.1969	20850	4	0.5142
	Mean	0.4356	0.4503	0.2004	20850	4.093	0.5314
	Variance	0.001282	0.001017	0.00135	55050	0.661	0.003973
FKMPP	Median	0.496	0.512	0.2671	21100	3	0.1938
	Mean	0.4952	0.5165	0.2706	21100	3.507	0.2087
	Variance	0.002498	0.001818	0.002036	57120	0.5245	0.002053

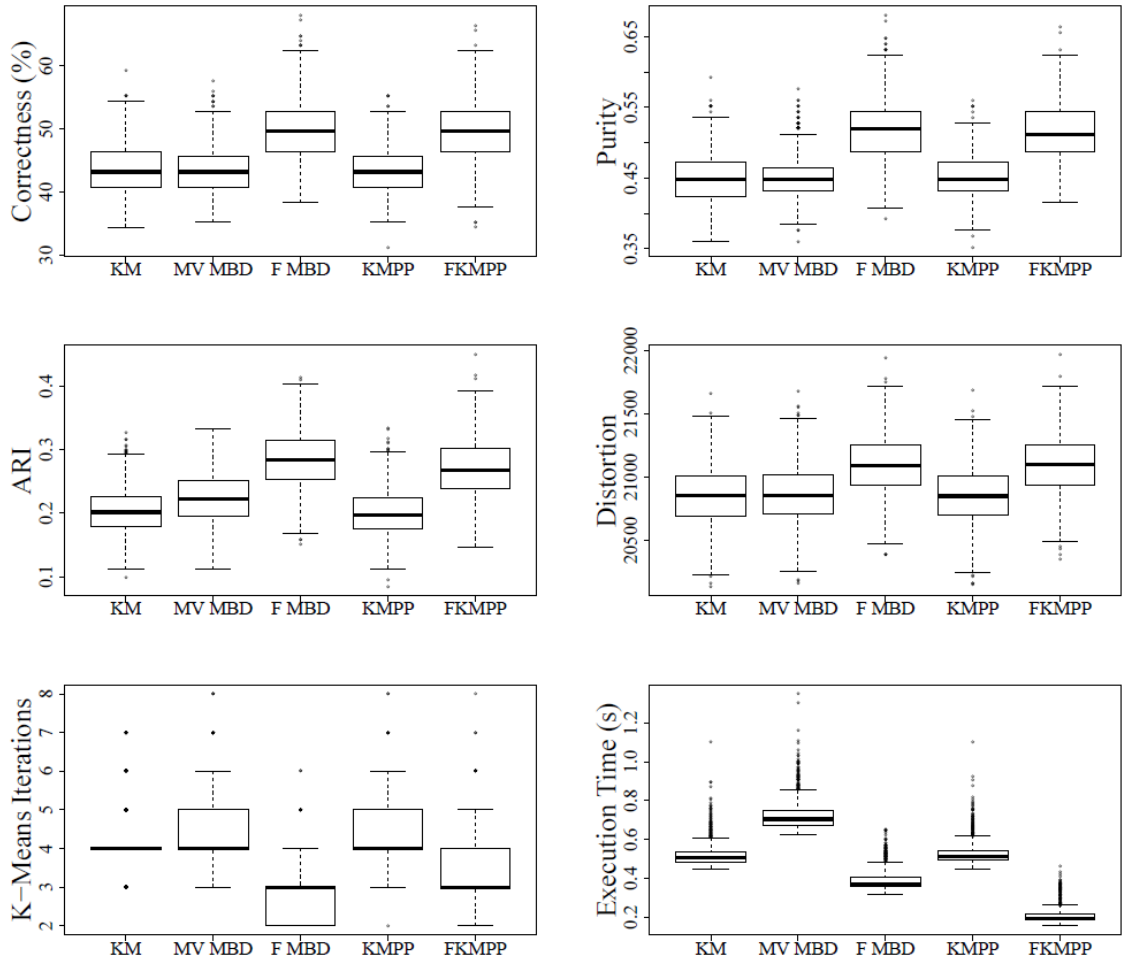


Fig. A.28. Model 3, 25% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.41.
SUMMARY STATISTICS FOR MODEL 3, 50% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.5, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.416	0.424	0.1552	18270	4	0.7103
	Mean	0.4157	0.4287	0.1584	18280	4.093	0.7353
	Variance	0.001287	0.001123	0.001282	68480	0.7091	0.006741
MV MBD	Median	0.408	0.424	0.1681	18280	4	0.9154
	Mean	0.4143	0.4289	0.1712	18280	4.136	0.9472
	Variance	0.001208	0.001089	0.001442	69940	0.9164	0.01069
FMBD	Median	0.448	0.472	0.2081	18520	3	0.3865
	Mean	0.4548	0.4694	0.211	18530	2.792	0.4043
	Variance	0.00163	0.001375	0.00157	72300	0.5713	0.003501
KMPP	Median	0.412	0.424	0.1559	18270	4	0.7184
	Mean	0.4155	0.4287	0.1569	18280	4.141	0.7435
	Variance	0.001343	0.00118	0.001294	68440	0.7178	0.006799
FKMPP	Median	0.448	0.464	0.2	18530	4	0.1983
	Mean	0.4503	0.4662	0.2029	18540	3.741	0.2095
	Variance	0.001735	0.00141	0.001398	71730	0.6966	0.001561

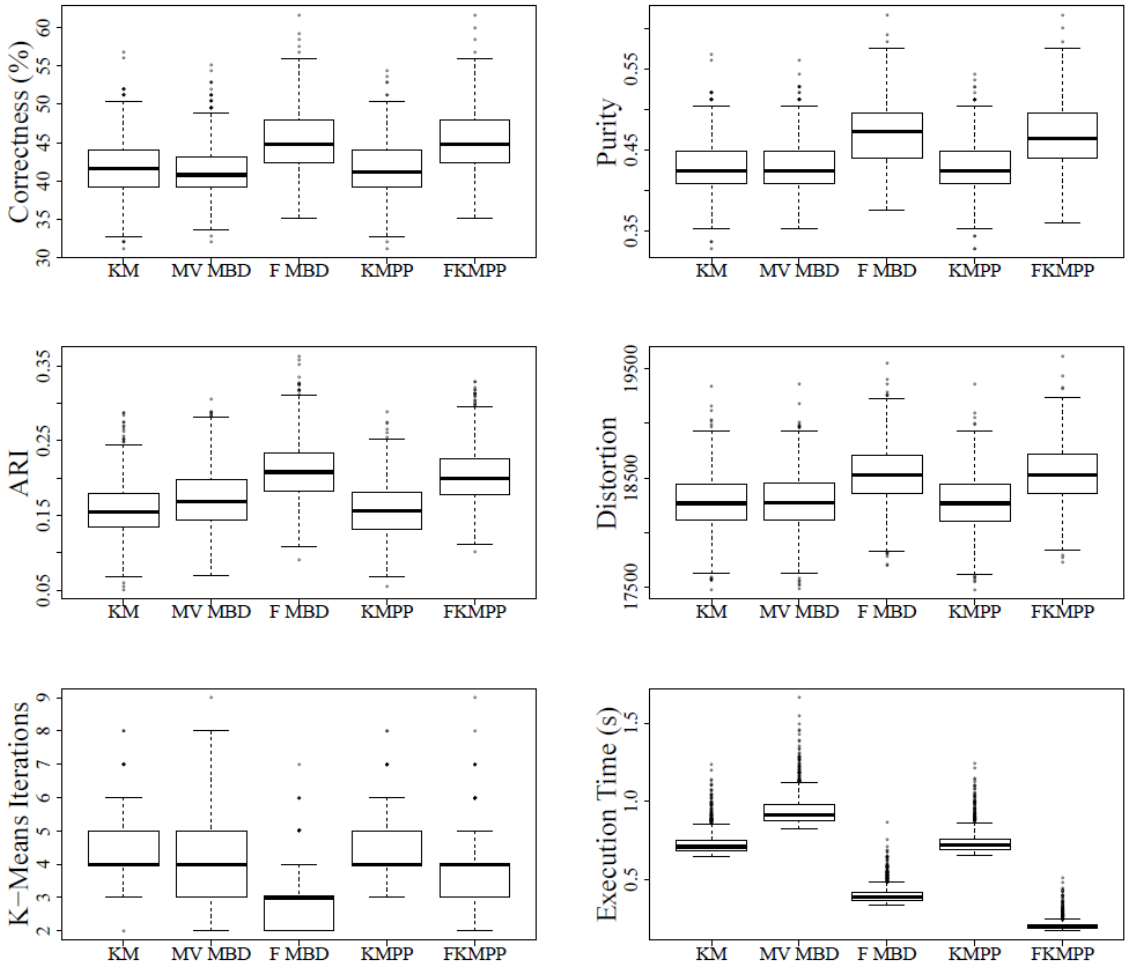


Fig. A.29. Model 3, 50% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.42.
SUMMARY STATISTICS FOR MODEL 3, 75% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.75, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.368	0.384	0.0883	15820	4	0.5211
	Mean	0.373	0.3843	0.08927	15810	4.12	0.5399
	Variance	0.001095	0.000942	0.000873	123500	0.7283	0.003457
MV MBD	Median	0.368	0.384	0.09226	15810	4	0.7095
	Mean	0.3736	0.3851	0.09393	15800	3.69	0.7336
	Variance	0.001054	0.000937	0.000919	122900	0.9048	0.006083
FMBD	Median	0.368	0.384	0.111	16370	3	0.3749
	Mean	0.3721	0.3821	0.1115	16370	2.842	0.3919
	Variance	0.001069	0.00103	0.001366	183900	0.6709	0.003257
KMPP	Median	0.368	0.384	0.08746	15830	4	0.5294
	Mean	0.3718	0.3837	0.08962	15810	4.066	0.5485
	Variance	0.001058	0.000974	0.000853	122200	0.7204	0.003589
FKMPP	Median	0.368	0.384	0.09828	16300	4	0.1901
	Mean	0.3731	0.3834	0.101	16300	3.819	0.2016
	Variance	0.001051	0.000973	0.001172	161600	0.6469	0.001471

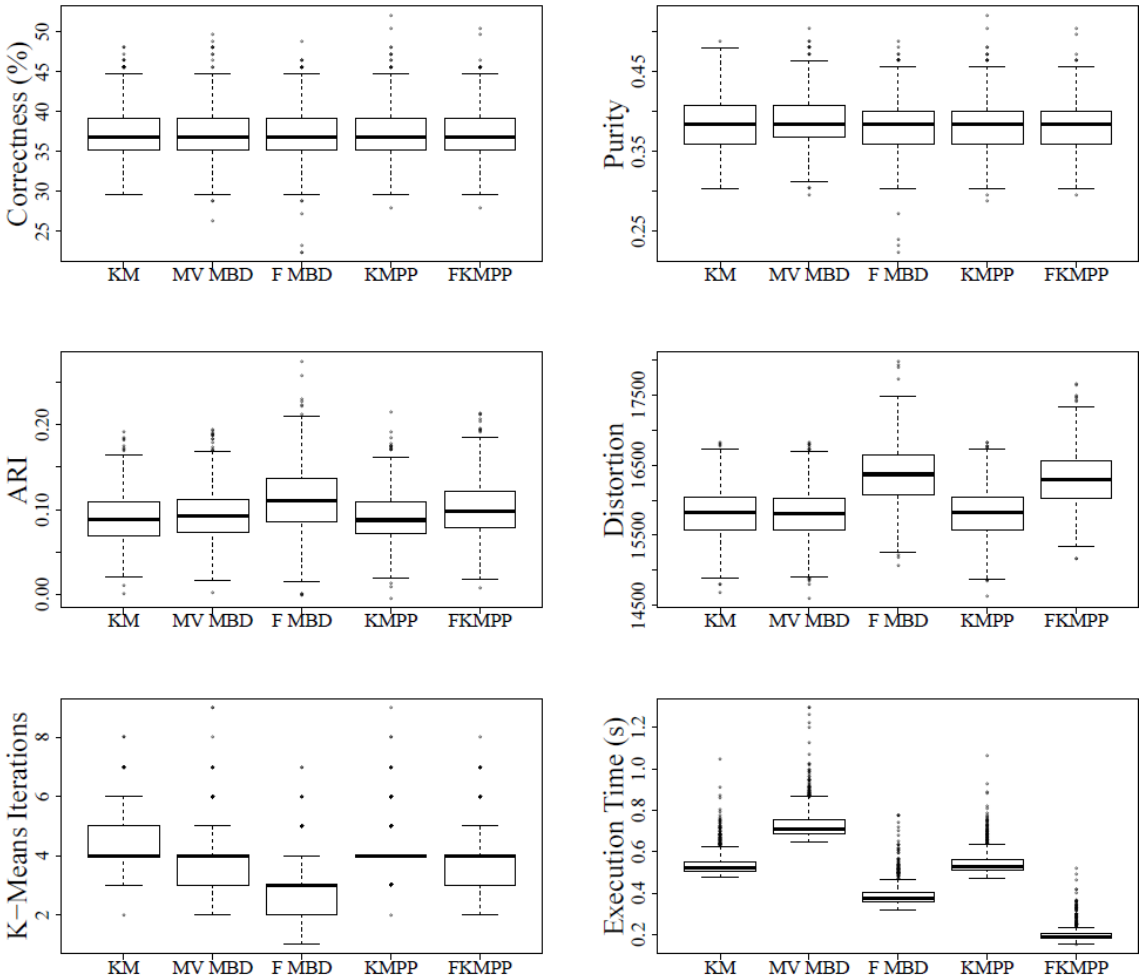


Fig. A.30. Model 3, 75% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

Model Four - Five-Way Comparison

TABLE A.43.
SUMMARY STATISTICS FOR MODEL 4, 5-WAY COPADIT FOR SIGMA = 0.

sigma = 0							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	1	1	1	0	1	~ 0
	Mean	1	1	1	0	1	4.695e-04
	Variance	0	0	0	0	0	2.127e-06
MV MBD	Median	1	1	1	0	1	2.666e-02
	Mean	1	1	1	0	1	2.772e-02
	Variance	0	0	0	0	0	8.619e-05
FMBD	Median	1	1	1	0	1	0.1285
	Mean	1	1	1	0	1	0.1351
	Variance	0	0	0	0	0	0.00062
KMPP	Median	1	1	1	0	1	1.661e-03
	Mean	1	1	1	0	1	2.322e-03
	Variance	0	0	0	0	0	8.763e-06
FKMPP	Median	1	1	1	0	1	0.1035
	Mean	1	1	1	0	1	0.1096
	Variance	0	0	0	0	0	0.0005184

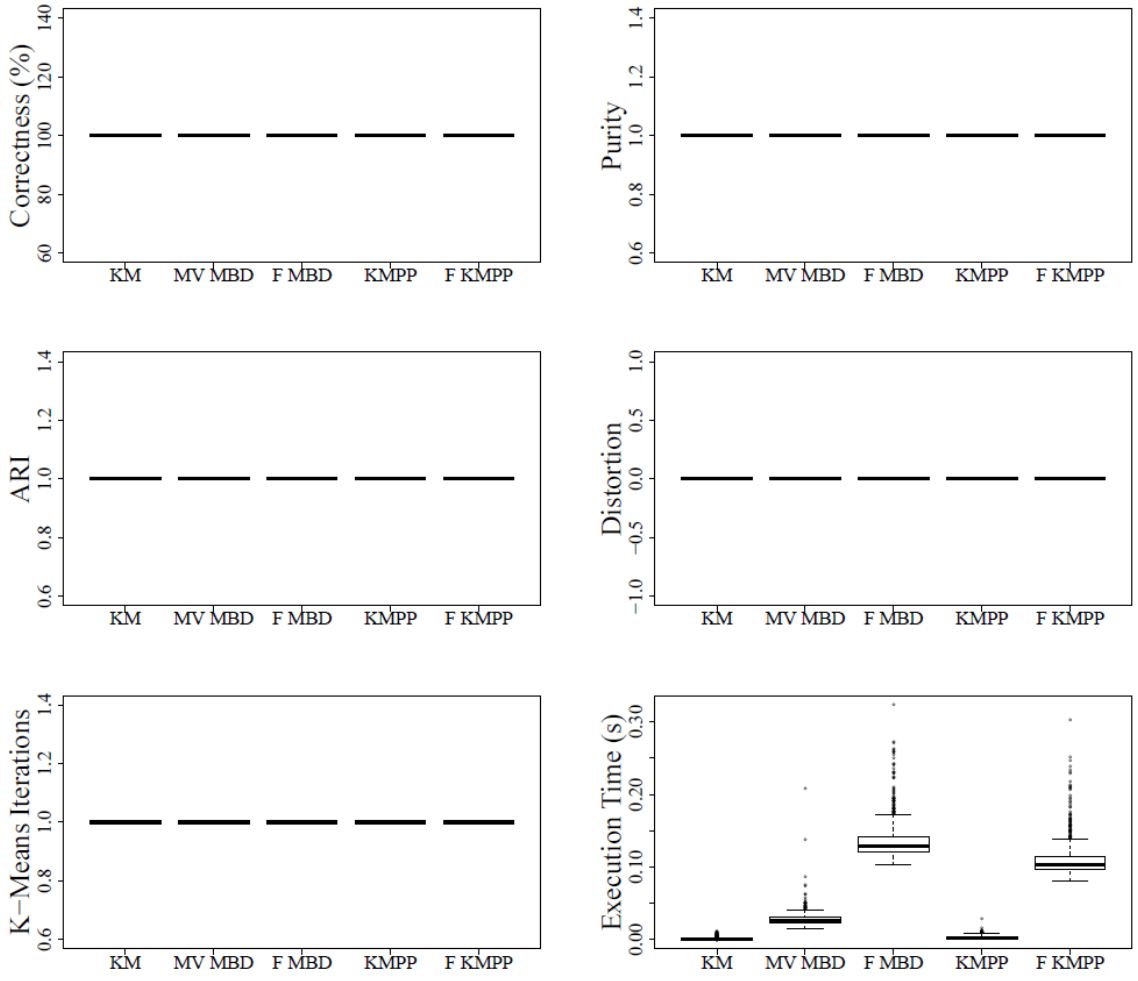


Fig. A.31. Model 4, 5-way CoPADIT measures distribution for sigma = 0.

TABLE A.44.
SUMMARY STATISTICS FOR MODEL 4, 5-WAY COPADIT FOR SIGMA = 0.5.

sigma = 0.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.72	0.74	0.5747	489.4	3	~ 0
	Mean	0.7	0.7357	0.5647	489.9	2.759	3.204e-04
	Variance	0.007037	0.002613	0.005667	329.2	0.3793	1.330e-06
MV MBD	Median	0.73	0.74	0.5802	488.6	2	2.195e-02
	Mean	0.7177	0.7447	0.5811	488.1	2.216	2.335e-02
	Variance	0.002986	0.000823	0.001962	259.6	0.2256	5.546e-05
FMBD	Median	0.76	0.76	0.6041	493	2	0.1207
	Mean	0.7518	0.7659	0.6031	492.9	1.962	0.1261
	Variance	0.003222	0.001368	0.001981	254.1	0.07463	0.000587
KMPP	Median	0.72	0.74	0.5768	488.8	3	1.556e-03
	Mean	0.702	0.7384	0.5671	489.6	2.671	2.212e-03
	Variance	0.006168	0.002131	0.004744	324.7	0.3751	1.148e-05
FKMPP	Median	0.73	0.75	0.5885	495.3	2	0.1004
	Mean	0.7141	0.7522	0.5852	495.3	2.309	0.1052
	Variance	0.006349	0.002093	0.003914	292.6	0.2478	0.0004778

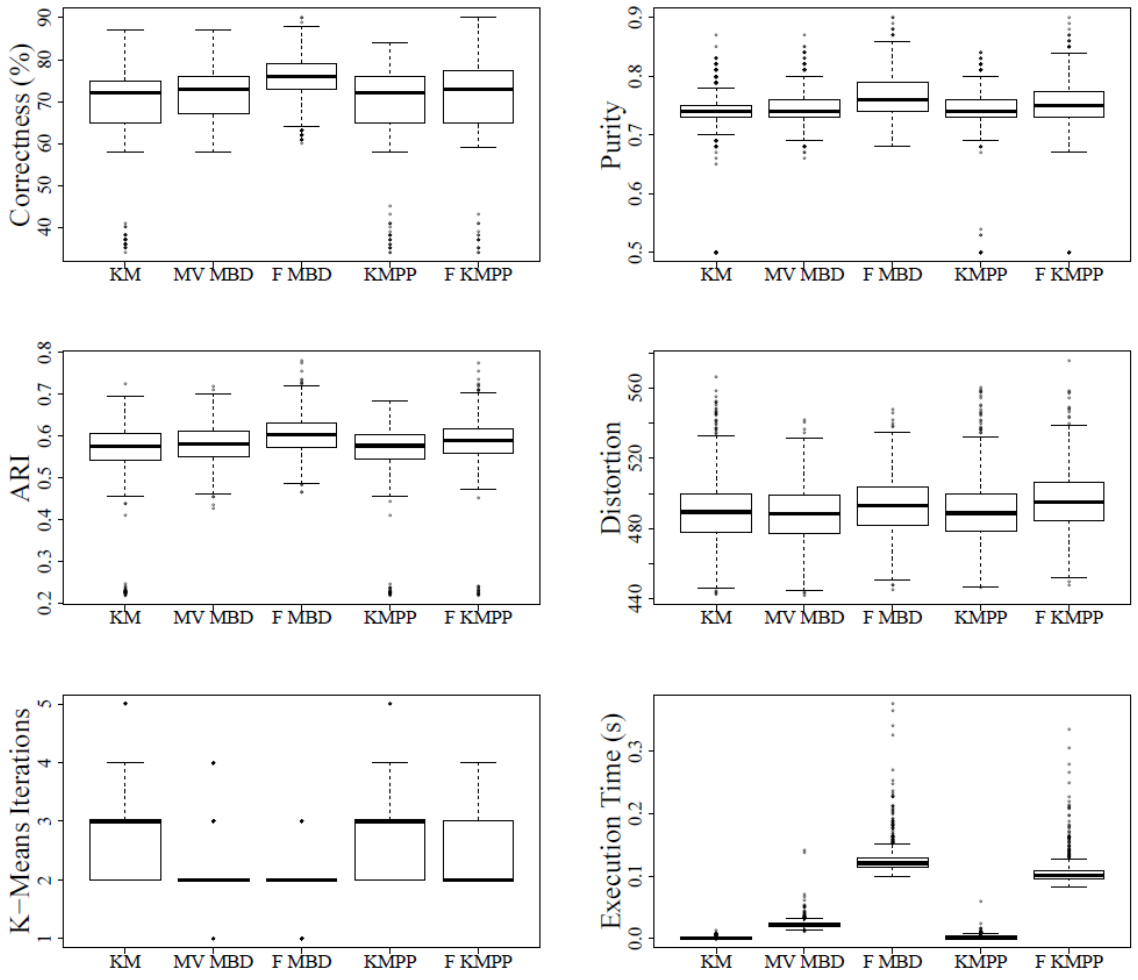


Fig. A.32. Model 4, 5-way CoPADIT measures distribution for sigma = 0.5.

TABLE A.45.
SUMMARY STATISTICS FOR MODEL 4, 5-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.57	0.59	0.307	1894	3	~ 0
	Mean	0.5692	0.5865	0.3071	1894	3.394	3.428e-04
	Variance	0.002879	0.001819	0.003109	3753	0.4652	1.338e-06
MV MBD	Median	0.58	0.6	0.3176	1891	3	0.02082
	Mean	0.5806	0.5955	0.3183	1891	2.667	0.02272
	Variance	0.002751	0.001852	0.003224	3743	0.4646	0.0000837
FMBD	Median	0.63	0.64	0.3737	1935	2	0.1128
	Mean	0.6293	0.6398	0.3772	1933	2.092	0.1178
	Variance	0.002007	0.001471	0.003132	4007	0.1357	0.0004105
KMPP	Median	0.58	0.59	0.3123	1895	3	1.559e-03
	Mean	0.5719	0.5886	0.3089	1894	3.364	2.060e-03
	Variance	0.003266	0.001969	0.003619	3695	0.512	7.644e-06
FKMPP	Median	0.62	0.64	0.3657	1940	3	0.09325
	Mean	0.611	0.6347	0.3643	1938	2.902	0.09774
	Variance	0.003055	0.001581	0.003792	3934	0.4789	0.0003549

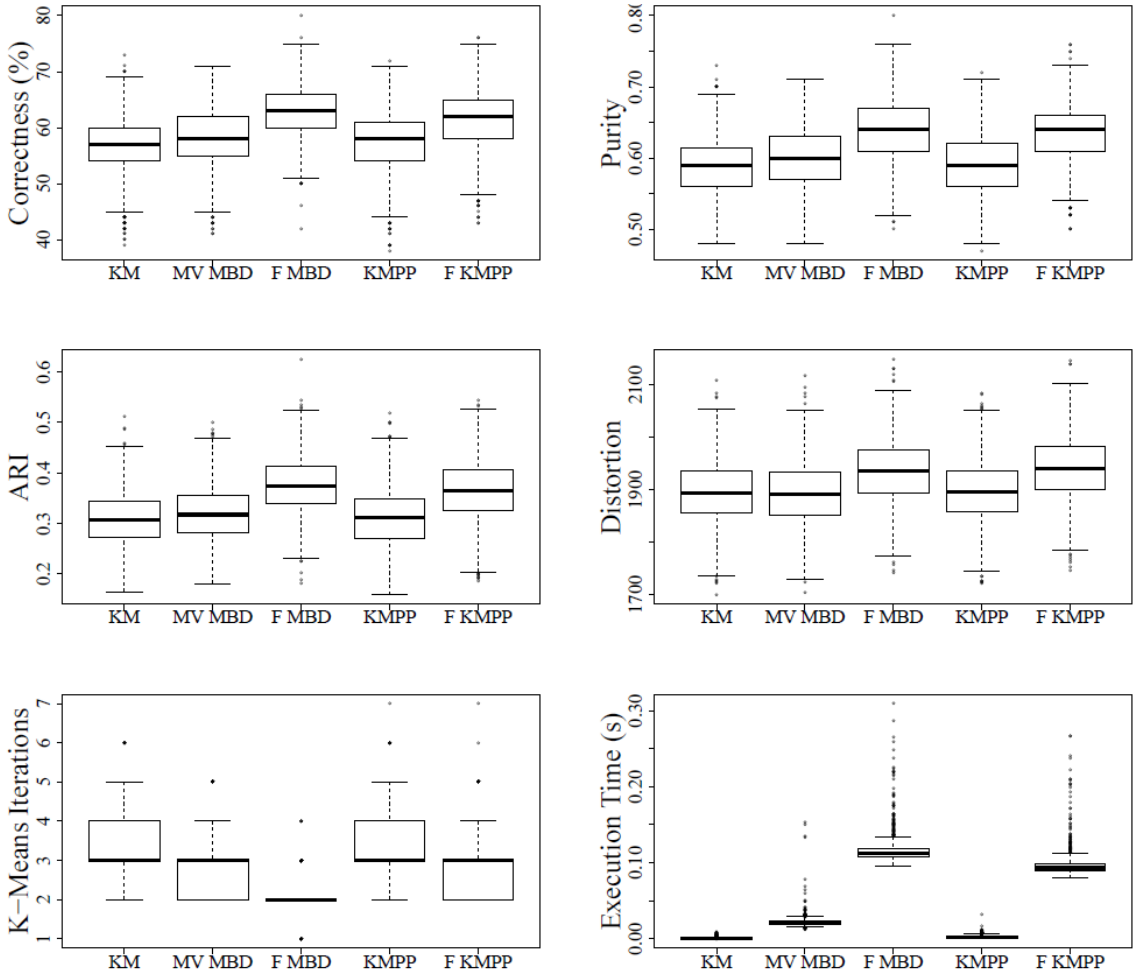


Fig. A.33. Model 4, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.46.
SUMMARY STATISTICS FOR MODEL 4, 5-WAY COPADIT FOR SIGMA = 1.5.

sigma = 1.5							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.48	0.49	0.1673	4165	3	~ 0
	Mean	0.479600	0.4968	0.1713	4164	3.568	3.910e-04
	Variance	0.002084	0.001375	0.002358	17670	0.4959	1.552e-06
MV MBD	Median	0.48	0.5	0.1685	4158	3	2.243e-02
	Mean	0.4807	0.4975	0.172	4158	2.852	2.363e-02
	Variance	0.001859	0.001314	0.002087	17640	0.5587	4.117e-05
FMBD	Median	0.53	0.54	0.2137	4294	2	0.1216
	Mean	0.5242	0.5371	0.2168	4291	2.201	0.1271
	Variance	0.002144	0.001602	0.002451	19380	0.2128	0.0005249
KMPP	Median	0.48	0.49	0.1662	4167	3	1.993e-03
	Mean	0.4776	0.4946	0.17	4165	3.55	2.391e-03
	Variance	0.00212	0.001426	0.002475	17740	0.538	8.769e-06
FKMPP	Median	0.52	0.53	0.2097	4300	3	0.1004
	Mean	0.518	0.5346	0.2133	4299	3.121	0.1054
	Variance	0.002462	0.001658	0.002544	19400	0.4949	0.0004482

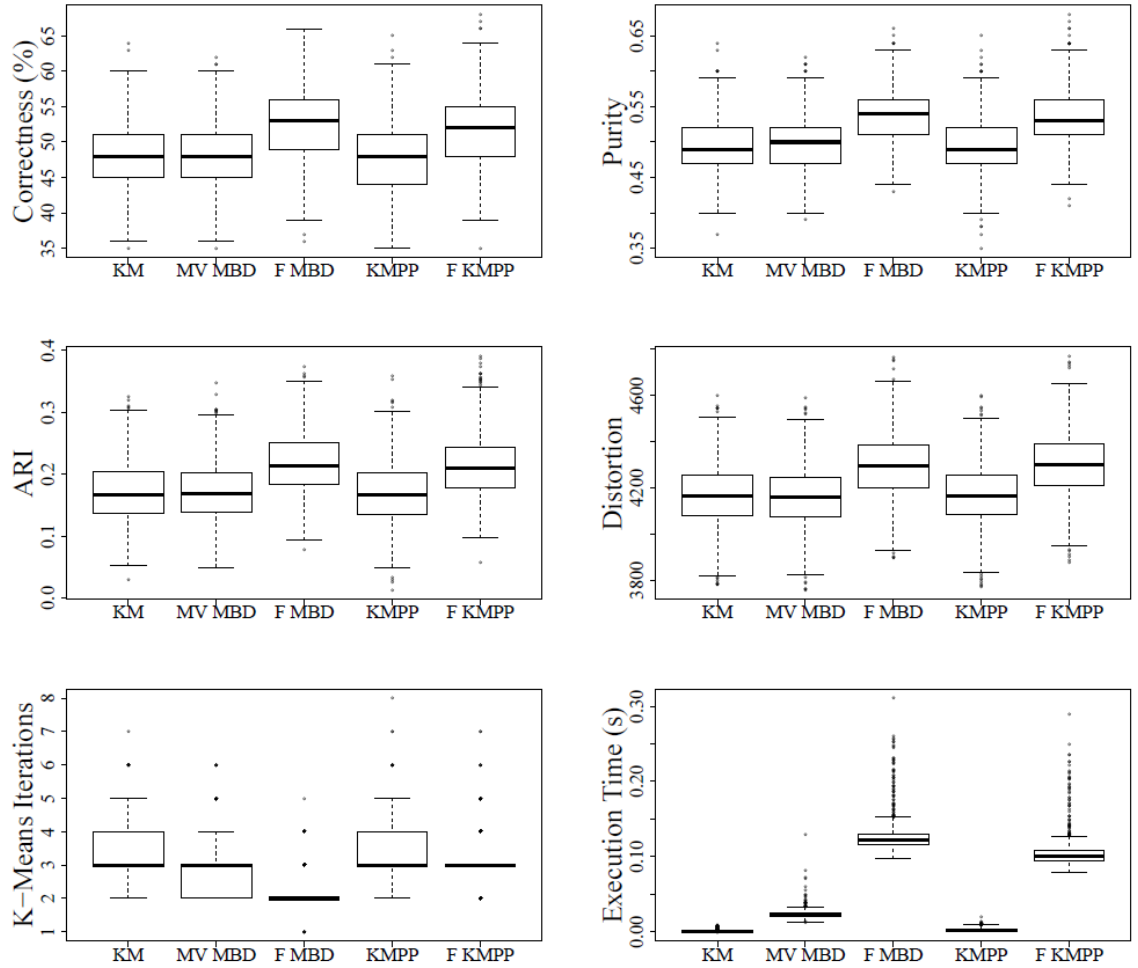


Fig. A.34. Model 4, 5-way CoPADIT measures distribution for sigma = 1.5.

TABLE A.47.
SUMMARY STATISTICS FOR MODEL 4, 5-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.42	0.44	0.08912	7295	4	~ 0
	Mean	0.4219	0.4381	0.09119	7293	3.681	5.878e-04
	Variance	0.001872	0.001631	0.001816	54510	0.5738	2.660e-06
MV MBD	Median	0.42	0.44	0.09033	7276	3	2.627e-02
	Mean	0.424	0.4402	0.09394	7280	2.997	2.729e-02
	Variance	0.001823	0.001488	0.001739	53230	0.5375	5.986e-05
FMBD	Median	0.46	0.48	0.1328	7566	2	0.1368
	Mean	0.4642	0.4785	0.1377	7560	2.188	0.1467
	Variance	0.001856	0.001399	0.001842	60190	0.1808	0.001144
KMPP	Median	0.42	0.44	0.08826	7296	4	2.056e-03
	Mean	0.4212	0.4371	0.0898	7294	3.676	2.652e-03
	Variance	0.002023	0.00168	0.001881	55330	0.6016	9.157e-06
FKMPP	Median	0.46	0.48	0.1313	7580	3	0.1131
	Mean	0.4594	0.4763	0.1341	7570	3.196	0.1221
	Variance	0.00191	0.001365	0.001797	60390	0.412	0.0008961

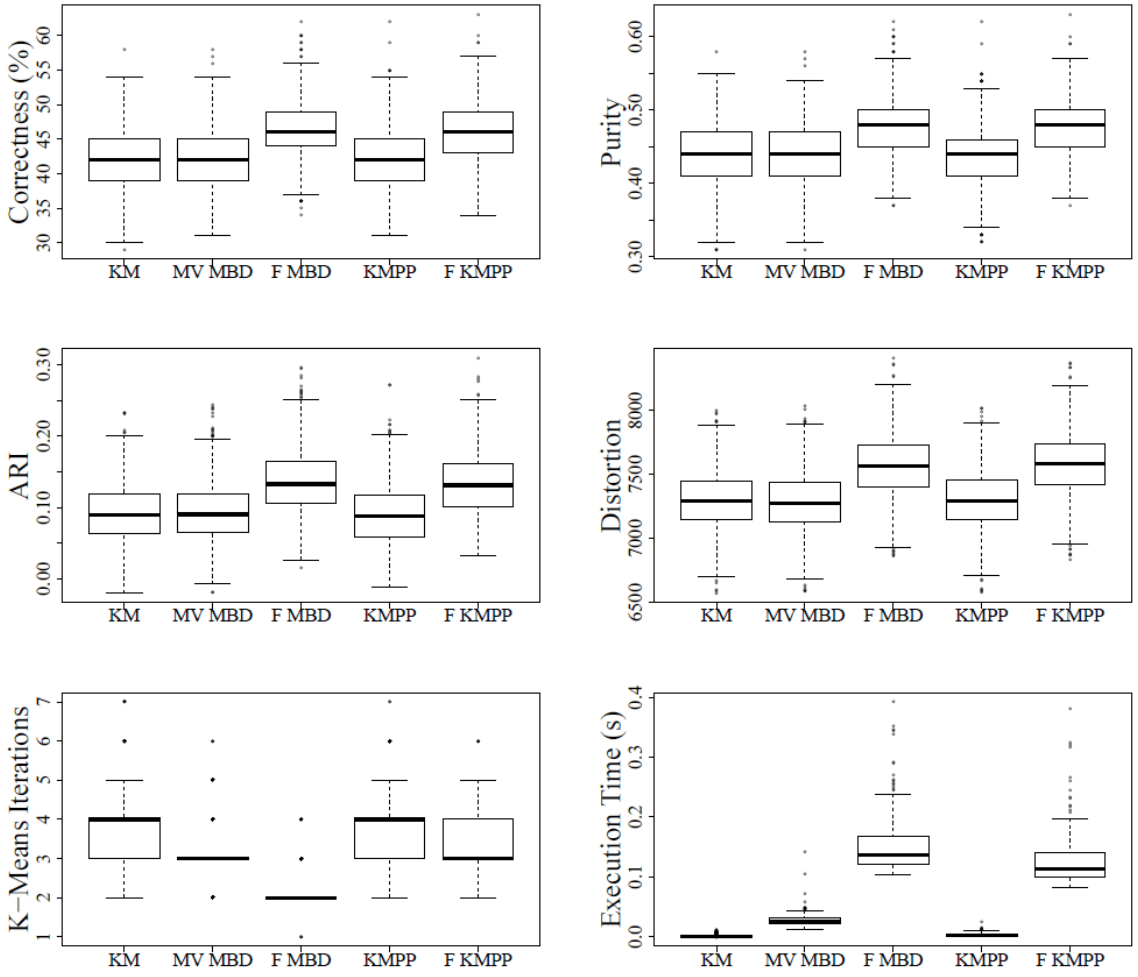


Fig. A.35. Model 4, 5-way CoPADIT measures distribution for sigma = 2.

The p-values for the paired t-test of correctness, purity and ARI for all methods are collected in the following tables for the different values of `sigma`.

TABLE A.48.
MODEL 4 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 0.5.

`sigma` = 0.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	2.150e-08	2.465e-54	6.387e-01	8.641e-05
	Purity		2.308e-07	8.075e-51	1.871e-01	3.221e-16
	ARI		1.346e-11	2.575e-47	4.116e-01	9.311e-14
MV MBD	Correctness	2.150e-08	-	2.550e-43	1.528e-07	2.316e-01
	Purity	2.308e-07		7.346e-53	1.200e-04	3.218e-06
	ARI	1.346e-11		1.043e-41	4.462e-10	5.249e-02
FMBD	Correctness	2.465e-54	2.550e-43	-	7.680e-55	3.931e-34
	Purity	8.075e-51	7.346e-53		5.969e-49	5.744e-17
	ARI	2.575e-47	1.043e-41		9.898e-49	8.470e-20
KMPP	Correctness	6.387e-01	1.528e-07	7.680e-55	-	4.647e-04
	Purity	1.871e-01	1.200e-04	5.969e-49		2.240e-12
	ARI	4.116e-01	4.462e-10	9.898e-49		1.112e-11
FKMPP	Correctness	8.641e-05	2.316e-01	3.931e-34	4.647e-04	-
	Purity	3.221e-16	3.218e-06	5.744e-17	2.240e-12	
	ARI	9.311e-14	5.249e-02	8.470e-20	1.112e-11	

TABLE A.49.
MODEL 4 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.

`sigma` = 1						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	8.819e-09	7.121e-144	2.180e-01	1.902e-68
	Purity		6.516e-10	9.160e-175	2.055e-01	1.118e-147
	ARI		4.773e-09	5.082e-175	4.058e-01	7.642e-113
MV MBD	Correctness	8.819e-09	-	1.862e-112	8.873e-06	4.561e-42
	Purity	6.516e-10		3.554e-140	1.129e-06	3.960e-113
	ARI	4.773e-09		9.662e-141	1.006e-06	7.727e-83
FMBD	Correctness	7.121e-144	1.862e-112	-	1.140e-127	6.603e-26
	Purity	9.160e-175	3.554e-140		3.311e-156	1.049e-06
	ARI	5.082e-175	9.662e-141		1.585e-155	4.526e-16
KMPP	Correctness	2.180e-01	8.873e-06	1.140e-127	-	9.396e-60
	Purity	2.055e-01	1.129e-06	3.311e-156		5.936e-131
	ARI	4.058e-01	1.006e-06	1.585e-155		3.352e-102
FKMPP	Correctness	1.902e-68	4.561e-42	6.603e-26	9.396e-60	-
	Purity	1.118e-147	3.960e-113	1.049e-06	5.936e-131	
	ARI	7.642e-113	7.727e-83	4.526e-16	3.352e-102	

TABLE A.50.
MODEL 4 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 1.5.

sigma = 1.5						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	5.013e-01	2.583e-106	2.420e-01	3.563e-75
	Purity		5.802e-01	1.959e-128	7.800e-02	4.792e-110
	ARI		6.503e-01	3.078e-116	4.171e-01	1.017e-96
MV MBD	Correctness	5.013e-01	-	1.909e-109	4.705e-02	9.523e-77
	Purity	5.802e-01		3.800e-133	1.284e-02	3.123e-113
	ARI	6.503e-01		1.592e-125	1.668e-01	8.112e-101
FMBD	Correctness	2.583e-106	1.909e-109	-	2.865e-116	2.339e-05
	Purity	1.959e-128	3.800e-133		2.376e-139	2.444e-02
	ARI	3.078e-116	1.592e-125		9.327e-121	9.995e-03
KMPP	Correctness	2.420e-01	4.705e-02	2.865e-116		2.374e-86
	Purity	7.800e-02	1.284e-02	2.376e-139		2.072e-122
	ARI	4.171e-01	1.668e-01	9.327e-121		1.970e-100
FKMPP	Correctness	3.563e-75	9.523e-77	2.339e-05	2.374e-86	-
	Purity	4.792e-110	3.123e-113	2.444e-02	2.072e-122	
	ARI	1.017e-96	8.112e-101	9.995e-03	1.970e-100	

TABLE A.51.
MODEL 4 ACCURACY MEASURES' P-VALUE OF THE PAIRED T-TEST FOR SIGMA = 2.

sigma = 2						
Method		KM	MV MBD	FMBD	KMPP	FKMPP
KM	Correctness	-	1.723e-01	7.508e-108	6.498e-01	4.626e-81
	Purity		1.164e-01	2.120e-124	4.559e-01	4.615e-107
	ARI		3.004e-02	2.632e-150	3.199e-01	2.494e-121
MV MBD	Correctness	1.723e-01	-	4.224e-107	5.386e-02	2.915e-80
	Purity	1.164e-01		3.831e-121	1.502e-02	4.100e-109
	ARI	3.004e-02		1.059e-143	9.019e-04	6.583e-118
FMBD	Correctness	7.508e-108	4.224e-107	-	3.397e-113	2.297e-04
	Purity	2.120e-124	3.831e-121		1.232e-128	3.109e-02
	ARI	2.632e-150	1.059e-143		4.741e-152	1.140e-03
KMPP	Correctness	6.498e-01	5.386e-02	3.397e-113	-	1.056e-90
	Purity	4.559e-01	1.502e-02	1.232e-128		6.170e-120
	ARI	3.199e-01	9.019e-04	4.741e-152		4.083e-133
FKMPP	Correctness	4.626e-81	2.915e-80	2.297e-04	1.056e-90	-
	Purity	4.615e-107	4.100e-109	3.109e-02	6.170e-120	
	ARI	2.494e-121	6.583e-118	1.140e-03	4.083e-133	

Model Four - Coefficient Clustering

TABLE A.52.
SUMMARY STATISTICS FOR MODEL 4, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 1.

sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.39	0.4	0.04746	2208	3	0.09115
	Mean	0.3918	0.403	0.05964	2201	2.86	0.09005
	Variance	0.002657	0.002565	0.002801	10170	0.4108	0.0002097
MV MBD	Median	0.38	0.4	0.044	2209	2	0.09656
	Mean	0.3913	0.402	0.05949	2203	2.019	0.1006
	Variance	0.003078	0.002975	0.003077	10550	0.09473	0.0003052
KMPP	Median	0.38	0.39	0.03823	2216	3	0.09169
	Mean	0.3853	0.3959	0.05258	2211	2.718	0.09144
	Variance	0.002736	0.002726	0.002814	9648	0.4069	0.0002386

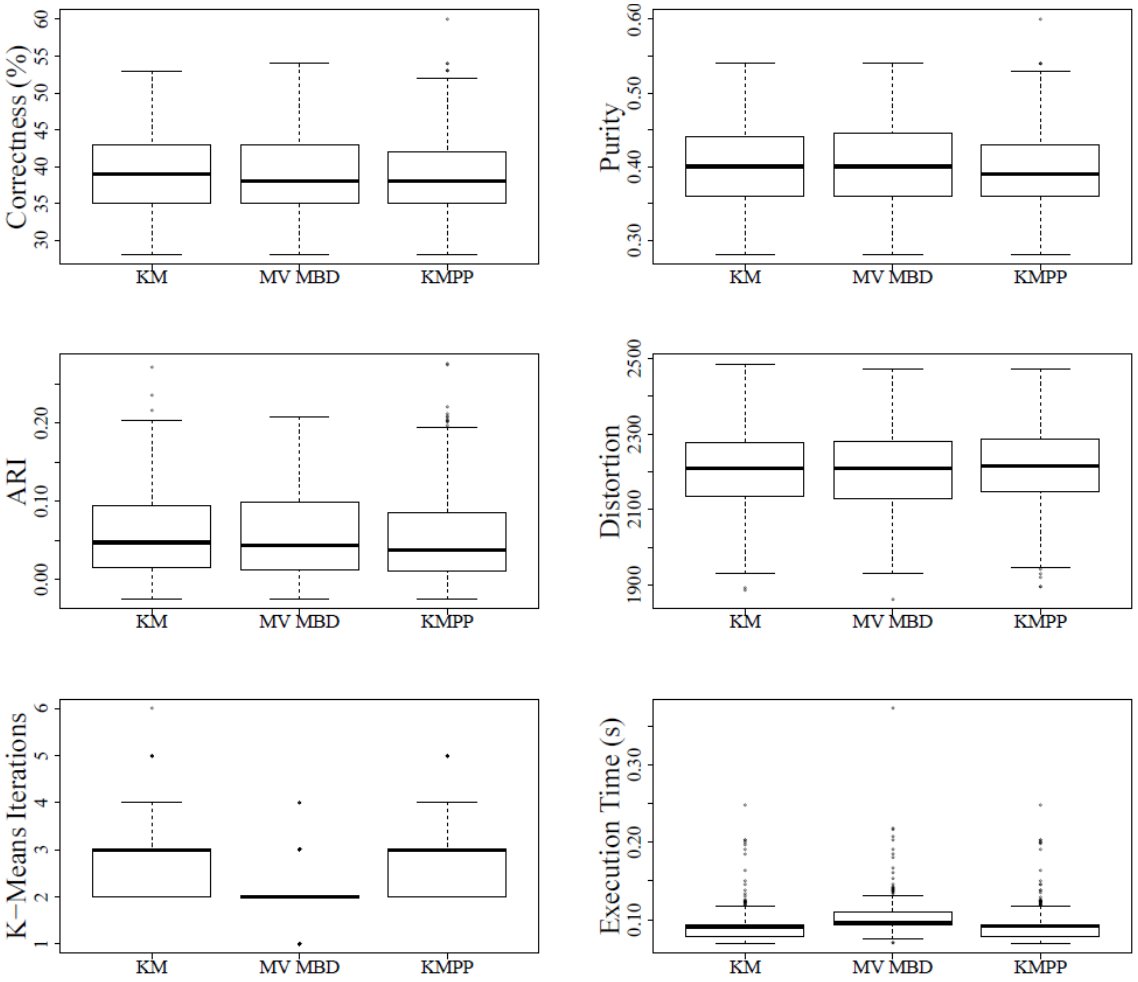


Fig. A.36. Model 4, 3-way CoPADIT measures distribution for sigma = 1.

TABLE A.53.
SUMMARY STATISTICS FOR MODEL 4, COEFFICIENT CLUSTERING 3-WAY COPADIT FOR SIGMA = 2.

sigma = 2							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.33	0.34	0.000692	8030	3	0.09077
	Mean	0.3367	0.3437	0.003683	8023	2.718	0.08923
	Variance	0.000777	0.00077	0.000258	68210	0.3909	0.0002324
MV MBD	Median	0.33	0.34	0.000606	8031	2	0.09377
	Mean	0.3364	0.344	0.004151	8035	1.973	0.09958
	Variance	0.0007415	0.00077	0.00026	68480	0.07435	0.0003144
KMPP	Median	0.33	0.34	0.000681	8038	2	0.09121
	Mean	0.3366	0.3437	0.00389	8032	1.973	0.09107
	Variance	0.0007844	0.000818	0.000278	69880	0.07435	0.0002613

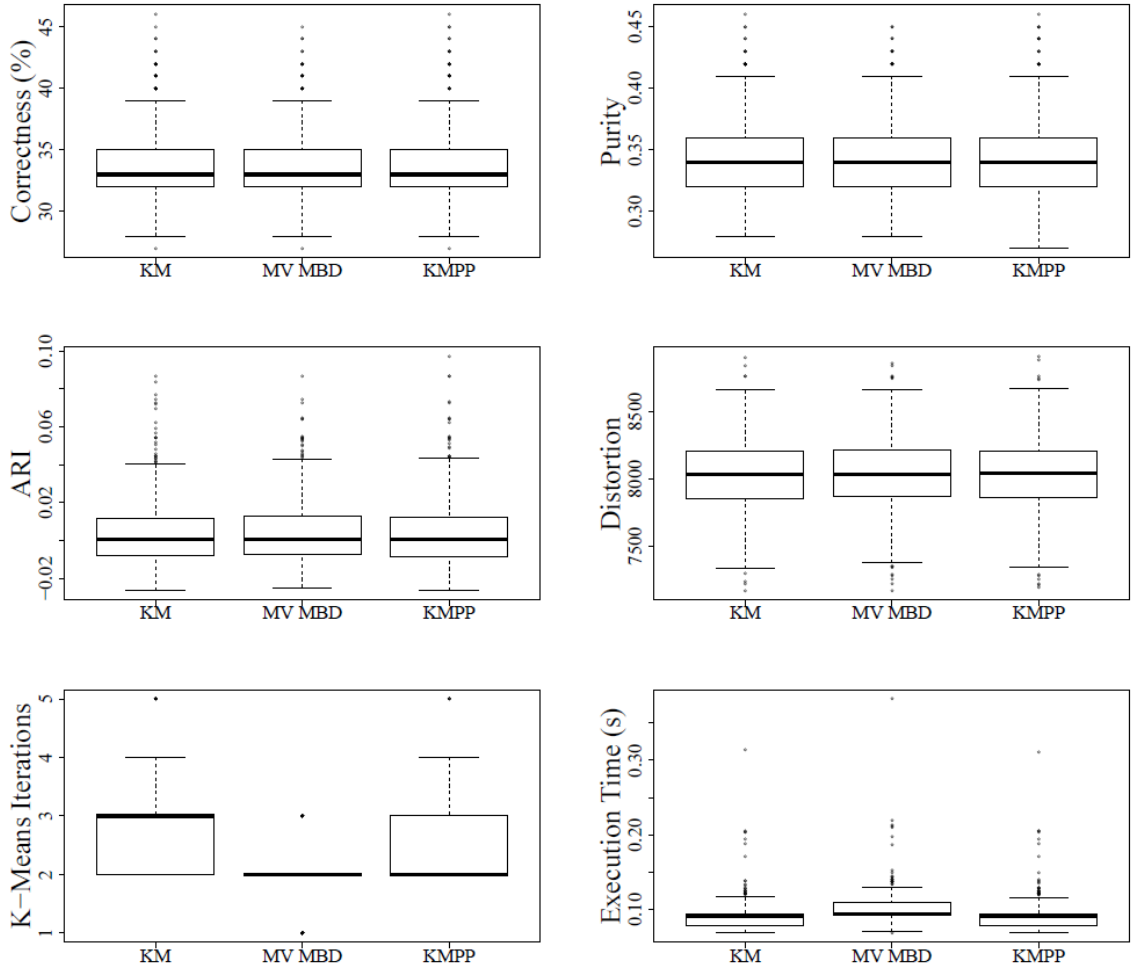


Fig. A.37. Model 4, 3-way CoPADIT measures distribution for sigma = 2.

Model Four - Missing Data

TABLE A.54.
SUMMARY STATISTICS FOR MODEL 4, 25% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.25, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.57	0.58	0.2858	1653	3	0.04323
	Mean	0.5634	0.5785	0.2843	1654	3.363	0.0405
	Variance	0.002783	0.001874	0.003211	3909	0.4737	0.0001433
MV MBD	Median	0.57	0.58	0.2896	1650	2	0.06248
	Mean	0.5691	0.583	0.2897	1651	2.54	0.0616
	Variance	0.002112	0.001546	0.002901	3832	0.4288	0.0001986
FMBD	Median	0.6	0.61	0.3238	1682	2	0.1094
	Mean	0.5959	0.607	0.3205	1683	2.119	0.1135
	Variance	0.00208	0.001551	0.003212	4030	0.149	0.0003599
KMPP	Median	0.56	0.58	0.2813	1654	3	0.04614
	Mean	0.5626	0.5783	0.2818	1655	3.329	0.04209
	Variance	0.002683	0.001798	0.003144	3844	0.4372	0.0001676
FKMPP	Median	0.58	0.6	0.3082	1687	3	0.09373
	Mean	0.5802	0.5995	0.3085	1688	2.981	0.09381
	Variance	0.003047	0.00187	0.003815	4091	0.4551	0.0002576

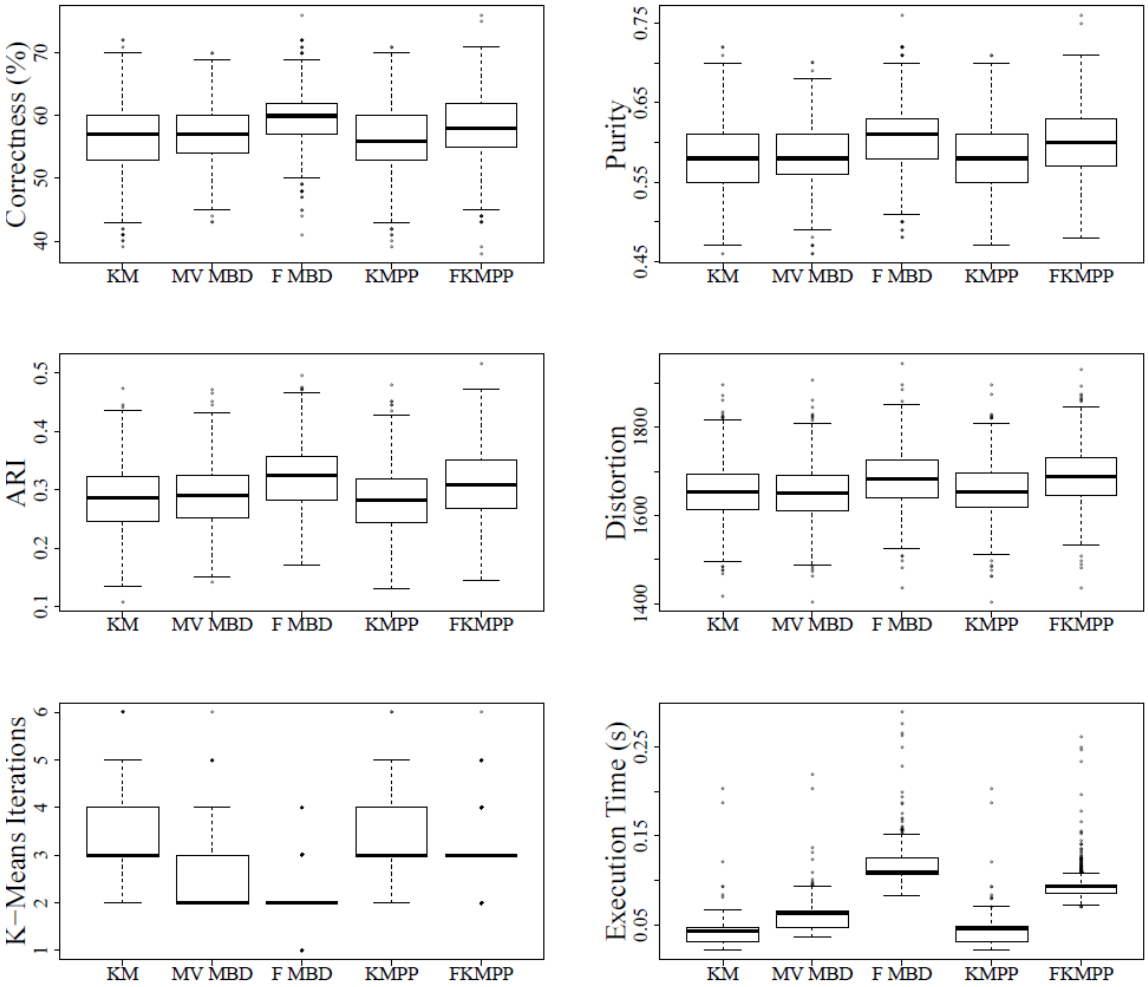


Fig. A.38. Model 4, 25% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.55.
SUMMARY STATISTICS FOR MODEL 4, 50% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.5, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.53	0.545	0.2244	1395	3	0.05982
	Mean	0.5296	0.5447	0.2265	1395	3.322	0.05813
	Variance	0.002335	0.001605	0.002767	4206	0.4948	0.0002168
MV MBD	Median	0.535	0.55	0.2264	1391	2	0.07811
	Mean	0.5353	0.5484	0.2315	1390	2.349	0.08046
	Variance	0.002242	0.001672	0.002861	4208	0.3536	0.0003203
FMBD	Median	0.54	0.56	0.2423	1434	2	0.1209
	Mean	0.5432	0.557	0.243	1433	2.128	0.1213
	Variance	0.002353	0.001714	0.002998	4708	0.1378	0.0005263
KMPP	Median	0.53	0.54	0.226	1398	3	0.06245
	Mean	0.5286	0.5447	0.2262	1396	3.25	0.05982
	Variance	0.00252	0.001751	0.002864	4263	0.474	0.0002398
FKMPP	Median	0.54	0.55	0.2322	1440	3	0.09377
	Mean	0.5337	0.5512	0.2341	1437	3.046	0.1011
	Variance	0.002693	0.001839	0.003193	4852	0.4844	0.0004397

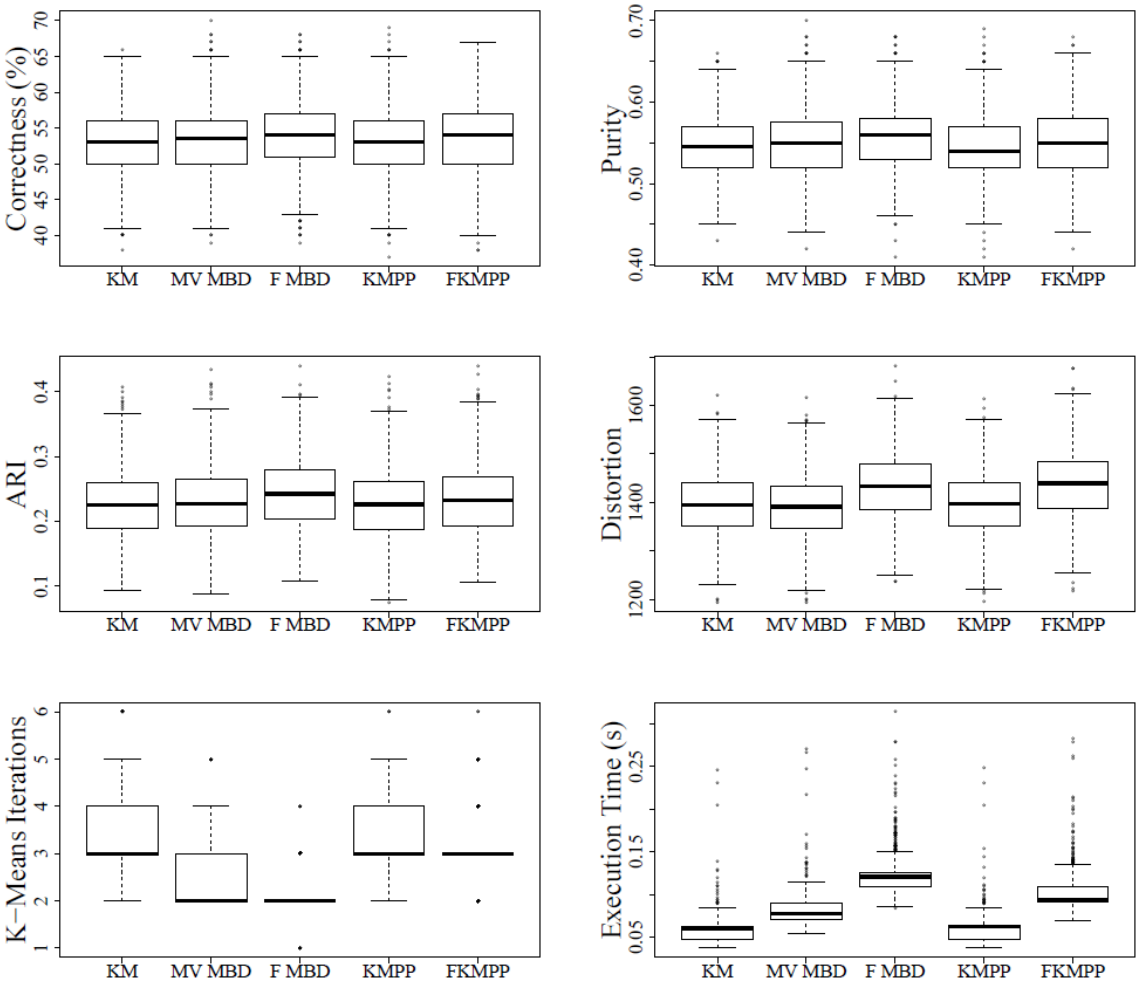


Fig. A.39. Model 4, 50% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

TABLE A.56.
SUMMARY STATISTICS FOR MODEL 4, 75% OF MISSING VALUES, 5-WAY COPADIT FOR SIGMA = 1.

Pmiss = 0.75, sigma = 1							
Method	Measure	Correctness	Purity	ARI	Distortion	Iterations	Exec. Time (seconds)
KM	Median	0.47	0.49	0.1441	1072	3	0.04686
	Mean	0.4733	0.4899	0.148	1073	3.179	0.04674
	Variance	0.001943	0.001429	0.0019	4838	0.4374	0.0001968
MV MBD	Median	0.48	0.49	0.1461	1064	2	0.06251
	Mean	0.4771	0.4919	0.1499	1066	2.162	0.06785
	Variance	0.002023	0.001522	0.002091	4700	0.1839	0.0002634
FMBD	Median	0.42	0.43	0.105	1383	2	0.1094
	Mean	0.4025	0.4132	0.09721	1428	1.852	0.1153
	Variance	0.004289	0.004378	0.003534	40690	0.2788	0.000475
KMPP	Median	0.47	0.49	0.1433	1072	3	0.04686
	Mean	0.4745	0.4906	0.1475	1073	3.12	0.04833
	Variance	0.002016	0.001461	0.00198	4792	0.432	0.0002321
FKMPP	Median	0.42	0.43	0.1062	1355	2	0.09373
	Mean	0.4128	0.4245	0.1013	1385	2.57	0.09537
	Variance	0.003126	0.003125	0.002851	31390	0.5416	0.0003253

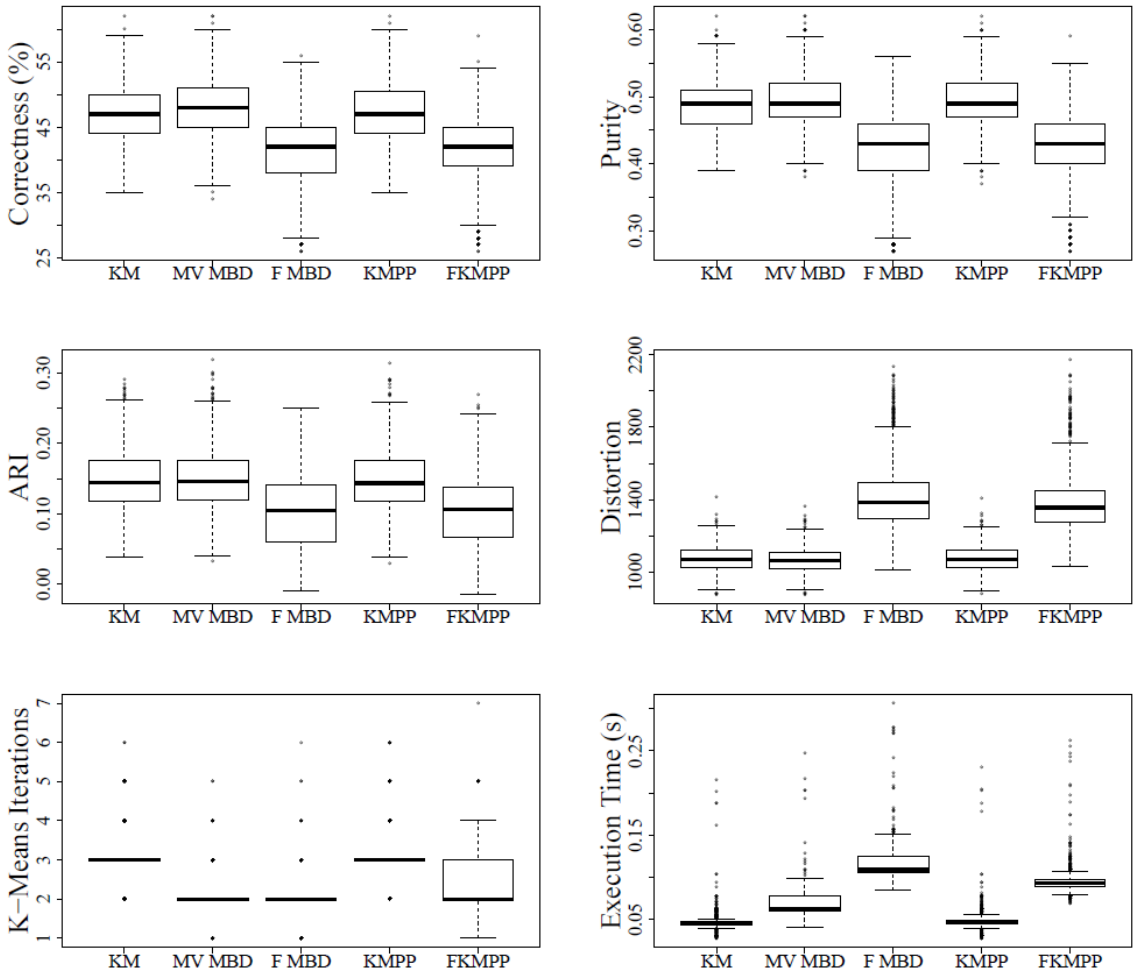


Fig. A.40. Model 4, 75% of missing values, 5-way CoPADIT measures distribution for sigma = 1.

Model One - OSF and DF Behavior According to Input Data (100 iterations)

TABLE A.57.
MEAN & VARIANCE FOR MODEL 1. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 0.5.

sigma = 0.5

			OSF			
			1	2	5	10
Correctness Mean	DF	4	0.9768	0.9767	0.9763	0.976
		6	0.9764	0.9757	0.9755	0.9753
		10	0.974	0.9741	0.9723	0.9718
		20	0.9566	0.9538	0.9527	0.9491
Correctness Variance						
	DF	4	0.0002826	0.0003052	0.0003023	0.000303
		6	0.0002556	0.0002753	0.0002674	0.0002938
		10	0.0003394	0.0002648	0.0002886	0.0002937
		20	0.004998	0.004793	0.005117	0.005885

			OSF			
			1	2	5	10
Purity Mean	DF	4	0.9768	0.9767	0.9763	0.976
		6	0.9764	0.9757	0.9755	0.9753
		10	0.974	0.9741	0.9723	0.9718
		20	0.9611	0.9578	0.9569	0.9541
Purity Variance						
	DF	4	0.0002826	0.0003052	0.0003023	0.000303
		6	0.0002556	0.0002753	0.0002674	0.0002938
		10	0.0003394	0.0002648	0.0002886	0.0002937
		20	0.002673	0.002785	0.00304	0.003445

			OSF			
			1	2	5	10
ARI Mean	DF	4	0.9406	0.9405	0.9395	0.9388
		6	0.9395	0.9378	0.9373	0.9369
		10	0.934	0.9338	0.9296	0.9284
		20	0.9145	0.9078	0.907	0.9027
ARI Variance						
	DF	4	0.001636	0.001766	0.00175	0.001755
		6	0.001527	0.00162	0.001568	0.001694
		10	0.001941	0.001564	0.001649	0.001665
		20	0.006028	0.006317	0.006272	0.006943

			OSF			
			1	2	5	10
Distortion Mean	DF	4	2420	2420	2420	2420
		6	2420	2420	2420	2420
		10	2420	2420	2420	2420
		20	2421	2421	2422	2422
Distortion Variance						
	DF	4	1335	1334	1334	1336
		6	1322	1320	1330	1326
		10	1335	1335	1322	1332
		20	1410	1457	1483	1483

			OSF			
			1	2	5	10
Iterations Mean	DF	4	1.33	1.38	1.37	1.38
		6	1.47	1.56	1.54	1.53
		10	1.69	1.7	1.7	1.72
		20	1.85	1.78	1.77	1.82
Iterations Variance						
	DF	4	0.2233	0.238	0.2355	0.238
		6	0.2516	0.2489	0.2509	0.2516
		10	0.2161	0.2121	0.2121	0.2036
		20	0.1288	0.1733	0.1789	0.1693

			OSF			
			1	2	5	10
Time Mean	DF	4	0.07263	0.1424	0.3454	0.7546
		6	0.06074	0.1247	0.2865	0.5939
		10	0.06518	0.1342	0.3094	0.6556
		20	0.0693	0.1326	0.3722	0.7238
Time Variance						
	DF	4	0.001879	0.007719	0.03756	0.3026
		6	0.0003887	0.008344	0.01009	0.0567
		10	0.0007387	0.008767	0.02105	0.2128
		20	0.002921	0.006858	0.0892	0.3536

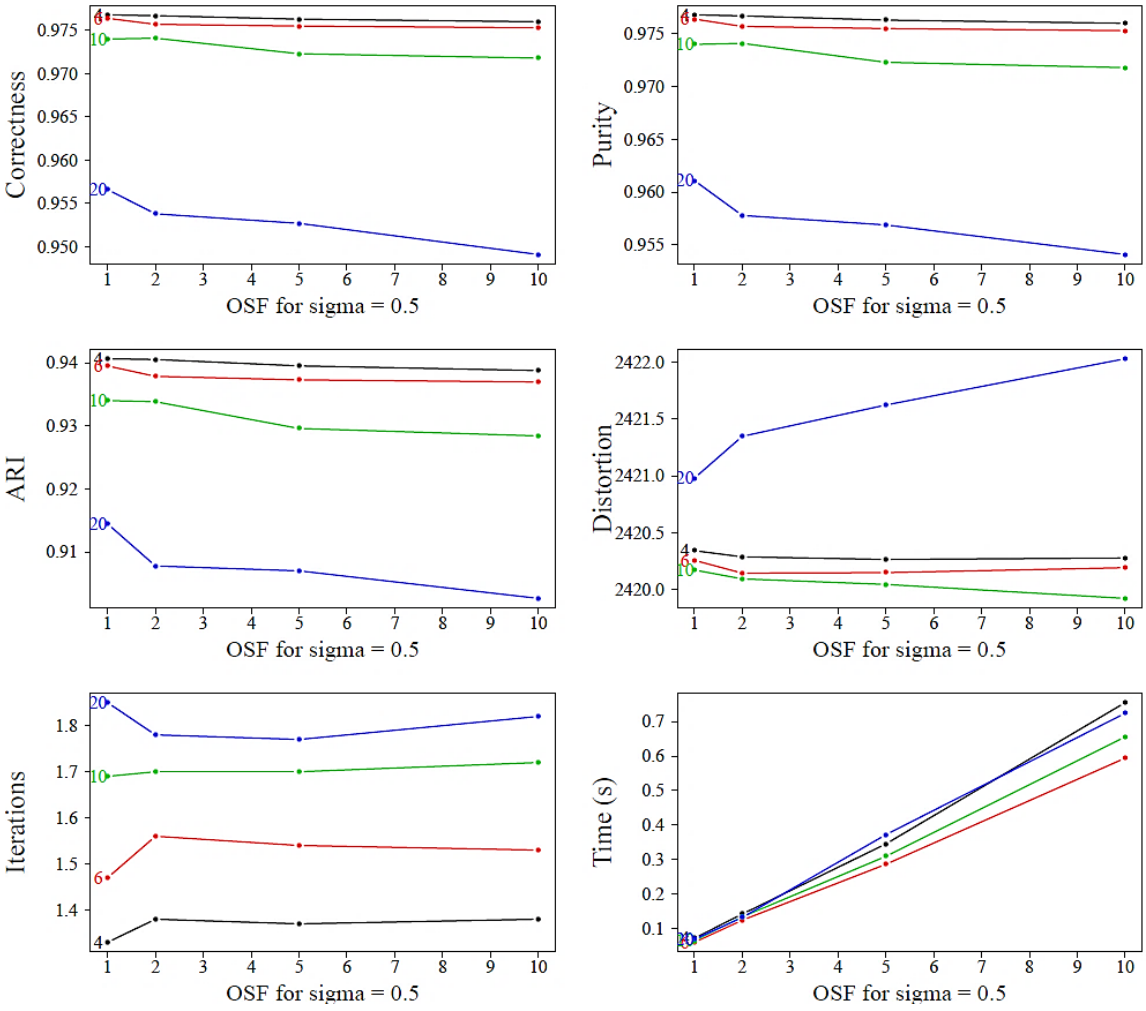


Fig. A.41. Model 1. OSF and DF behavior according to `sigma` = 0.5.

TABLE A.58.
MEAN AND VARIANCE FOR MODEL 1. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 1.

`sigma` = 1

			OSF			
			1	2	5	10
Correctness Mean	DF	4	0.8335	0.8264	0.8295	0.8283
		6	0.8129	0.8148	0.8126	0.8165
		10	0.7879	0.7886	0.7889	0.791
		20	0.7536	0.7553	0.7499	0.7566
Correctness Variance						
	DF	4	0.003791	0.004169	0.003473	0.004166
		6	0.004857	0.004755	0.004791	0.003983
		10	0.005968	0.005531	0.005614	0.006122
		20	0.004949	0.004682	0.004033	0.003544

			OSF			
			1	2	5	10
Purity Mean	DF	4	0.8357	0.8294	0.8315	0.8318
		6	0.816	0.8192	0.8167	0.8186
		10	0.7955	0.7956	0.7955	0.7974
		20	0.7641	0.7645	0.76	0.7627
Purity Variance						
	DF	4	0.003003	0.003151	0.002837	0.003031
		6	0.003887	0.003401	0.003441	0.00338
		10	0.003902	0.003774	0.003942	0.004397
		20	0.003067	0.003164	0.002285	0.002454

			OSF			
			1	2	5	10
ARI Mean	DF	4	0.6544	0.6466	0.6485	0.6498
		6	0.6303	0.6323	0.6297	0.6333
		10	0.6055	0.6063	0.6068	0.6096
		20	0.5716	0.5691	0.5645	0.5681
ARI Variance						
	DF	4	0.006227	0.006277	0.006278	0.006433
		6	0.007338	0.006917	0.006594	0.006802
		10	0.007205	0.006781	0.006438	0.007496
		20	0.004303	0.004638	0.00344	0.003872

			OSF			
			1	2	5	10
Distortion Mean	DF	4	9648	9647	9647	9647
		6	9643	9643	9643	9643
		10	9640	9639	9642	9640
		20	9633	9630	9633	9633
Distortion Variance						
	DF	4	20770	20820	20650	20530
		6	20400	20360	20700	20660
		10	20620	19960	21040	20380
		20	21130	22290	21540	20500

			OSF			
			1	2	5	10
Iterations Mean	DF	4	1.84	1.86	1.89	1.86
		6	2.04	2.03	1.97	1.96
		10	2.1	2.15	2.04	2.08
		20	2.41	2.21	2.22	2.21
Iterations Variance						
	DF	4	0.1358	0.1418	0.1393	0.1418
		6	0.1196	0.0496	0.09	0.1196
		10	0.1111	0.1894	0.1802	0.1349
		20	0.406	0.1878	0.2137	0.208

			OSF			
			1	2	5	10
Time Mean	DF	4	0.1372	0.2938	0.6568	1.459
		6	0.1391	0.2499	0.6241	1.178
		10	0.1604	0.29	0.7883	1.616
		20	0.1985	0.3876	0.9501	1.868
Time Variance						
	DF	4	0.01816	0.09592	0.3432	1.868
		6	0.01654	0.05079	0.3059	1.13
		10	0.01691	0.0534	0.4162	1.758
		20	0.0226	0.07776	0.5361	1.879

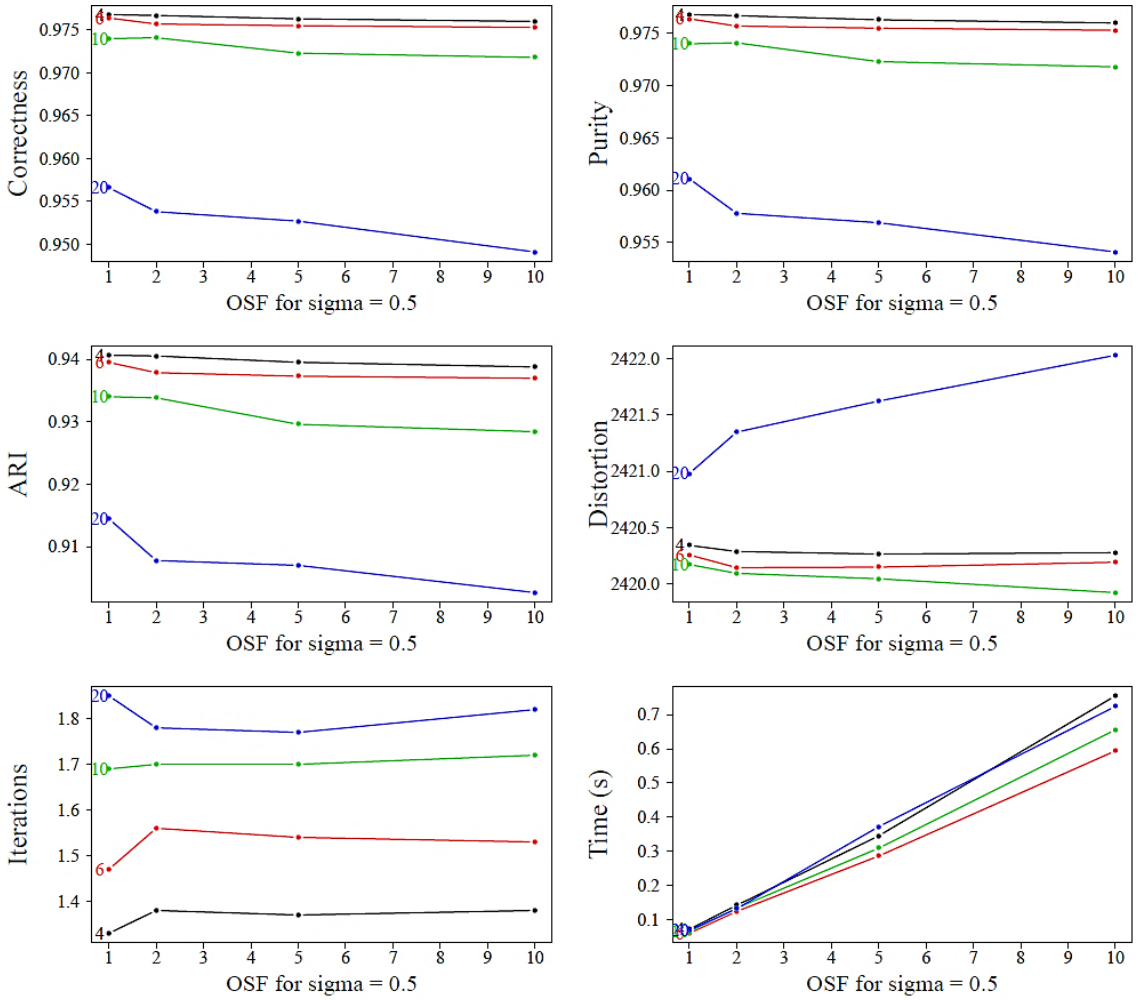


Fig. A.42. Model 1. OSF and DF behavior according to `sigma` = 1.

TABLE A.59.
MEAN & VARIANCE FOR MODEL 1. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 1.5.

`sigma` = 1.5

			OSF			
			1	2	5	10
Correctness Mean	DF	4	0,6626	0,659	0,6641	0,659
		6	0,649	0,6572	0,6503	0,6449
		10	0,6297	0,6327	0,6352	0,6385
		20	0,5958	0,5986	0,5991	0,6026
Correctness Variance						
	DF	4	0.004036	0.004132	0.004	0.003928
		6	0.003963	0.004014	0.004524	0.004353
		10	0.002829	0.004078	0.003207	0.003215
		20	0.003093	0.003244	0.003297	0.003335

			OSF			
			1	2	5	10
Purity Mean	DF	4	0.6702	0.6696	0.6728	0.6689
		6	0.6601	0.6652	0.661	0.6568
		10	0.6406	0.644	0.645	0.6476
		20	0.608	0.6086	0.6097	0.61
Purity Variance						
	DF	4	0.002774	0.002656	0.002608	0.002456
		6	0.002625	0.002799	0.002957	0.00279
		10	0.00181	0.002723	0.002072	0.002283
		20	0.002244	0.002289	0.00252	0.002582

			OSF			
			1	2	5	10
ARI Mean	DF	4	0.3797	0.3828	0.3816	0.3776
		6	0.3729	0.3774	0.3727	0.3698
		10	0.3492	0.3525	0.3553	0.3589
		20	0.3037	0.3062	0.3085	0.3075
ARI Variance						
	DF	4	0.003976	0.003912	0.0037	0.003483
		6	0.004114	0.004373	0.004663	0.004721
		10	0.00327	0.004613	0.003854	0.003794
		20	0.004234	0.004445	0.004361	0.004907

			OSF			
			1	2	5	10
Distortion Mean	DF	4	21630	21630	21630	21630
		6	21610	21610	21620	21610
		10	21590	21590	21590	21590
		20	21540	21540	21550	21540
Distortion Variance						
	DF	4	108600	109200	109000	108800
		6	110400	109100	108600	109400
		10	106300	110300	106700	108300
		20	114500	106100	107500	111000

			OSF			
			1	2	5	10
Iterations Mean	DF	4	2.06	2.09	2.13	2.15
		6	2.12	2.16	2.21	2.16
		10	2.27	2.27	2.33	2.27
		20	2.49	2.58	2.61	2.77
Iterations Variance						
	DF	4	0.1176	0.1635	0.1546	0.2096
		6	0.1471	0.156	0.2484	0.156
		10	0.1991	0.2395	0.2839	0.3203
		20	0.3332	0.3471	0.3817	0.5627

			OSF			
			1	2	5	10
Time Mean	DF	4	0.1837	0.3885	0.8333	1.68
		6	0.1921	0.3089	0.78	1.475
		10	0.1899	0.4395	0.9682	2.069
		20	0.2336	0.4096	1.07	2.055
Time Variance						
	DF	4	0.0269	0.103	0.5389	2.005
		6	0.02531	0.06015	0.3961	1.486
		10	0.01899	0.1019	0.4764	1.873
		20	0.02567	0.07931	0.5366	1.65

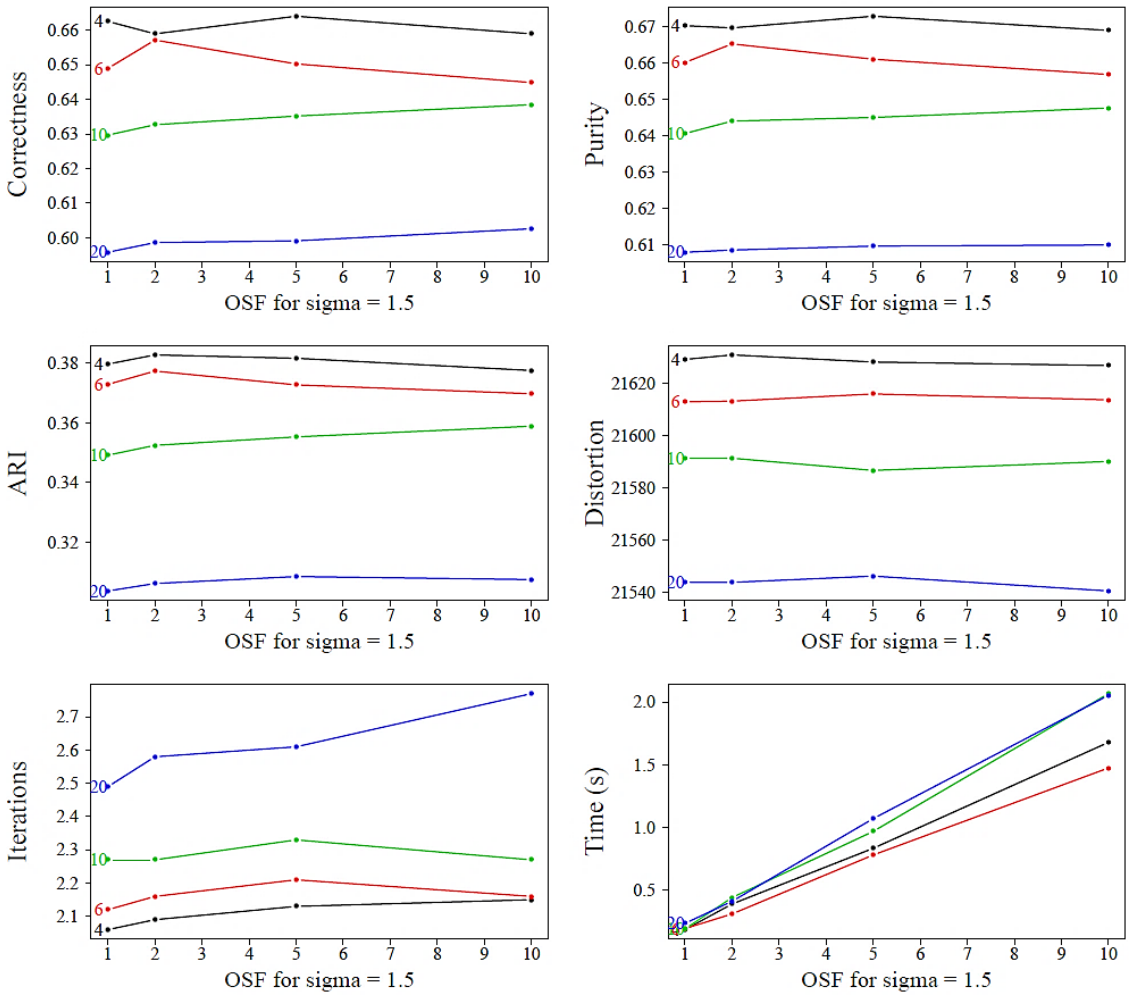


Fig. A.43. Model 1. OSF and DF behavior according to `sigma` = 1.5.

TABLE A.60.
MEAN AND VARIANCE FOR MODEL 1. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 2.

sigma = 2

			OSF			
			1	2	5	10
Correctness Mean	DF	4	0.5522	0.5563	0.556	0.5599
		6	0.5357	0.5399	0.5404	0.5411
		10	0.5191	0.5169	0.5226	0.5191
		20	0.4704	0.4821	0.4714	0.4722
Correctness Variance						
	DF	4	0.00347	0.003561	0.003653	0.003373
		6	0.004065	0.003682	0.004279	0.004254
		10	0.003515	0.003119	0.003025	0.003158
		20	0.003299	0.003685	0.003358	0.003102

			OSF			
			1	2	5	10
Purity Mean	DF	4	0.5626	0.5665	0.5656	0.5686
		6	0.548	0.554	0.5532	0.5549
		10	0.5313	0.5322	0.5353	0.5315
		20	0.485	0.4962	0.4844	0.4873
Purity Variance						
	DF	4	0.002567	0.002625	0.002827	0.002545
		6	0.002818	0.002477	0.00307	0.003047
		10	0.002615	0.002341	0.002407	0.002637
		20	0.002652	0.003095	0.002896	0.002499

			OSF			
			1	2	5	10
ARI Mean	DF	4	0.2239	0.2272	0.2274	0.2294
		6	0.2073	0.2153	0.2132	0.2148
		10	0.1893	0.1907	0.1948	0.1926
		20	0.1331	0.1478	0.136	0.1397
ARI Variance						
	DF	4	0.003383	0.00343	0.003724	0.003607
		6	0.004051	0.003373	0.003623	0.004003
		10	0.003532	0.003285	0.003159	0.003884
		20	0.003088	0.003474	0.003213	0.003465

			OSF			
			1	2	5	10
Distortion Mean	DF	4	38340	38340	38350	38340
		6	38300	38290	38300	38300
		10	38240	38250	38240	38250
		20	38140	38130	38120	38130
Distortion Variance						
	DF	4	359300	360700	373000	358900
		6	358300	357600	361900	360800
		10	374800	361500	360600	365000
		20	385700	355100	365600	360700

			OSF			
			1	2	5	10
Iterations Mean	DF	4	2.2	2.18	2.22	2.17
		6	2.27	2.3	2.24	2.3
		10	2.59	2.48	2.48	2.52
		20	2.89	2.88	2.81	2.89
Iterations Variance						
	DF	4	0.2222	0.2097	0.1935	0.1627
		6	0.2597	0.2929	0.2651	0.2727
		10	0.608	0.4339	0.4339	0.4137
		20	0.4625	0.4703	0.4787	0.4625

			OSF			
			1	2	5	10
Time Mean	DF	4	0.2085	0.381	0.8753	1.677
		6	0.1798	0.4061	0.9183	1.786
		10	0.2244	0.4446	1.214	1.942
		20	0.2462	0.4984	1.219	2.341
Time Variance						
	DF	4	0.03122	0.108	0.5134	1.867
		6	0.01862	0.09428	0.4547	1.721
		10	0.02534	0.09376	0.572	1.979
		20	0.02807	0.09013	0.5951	1.914

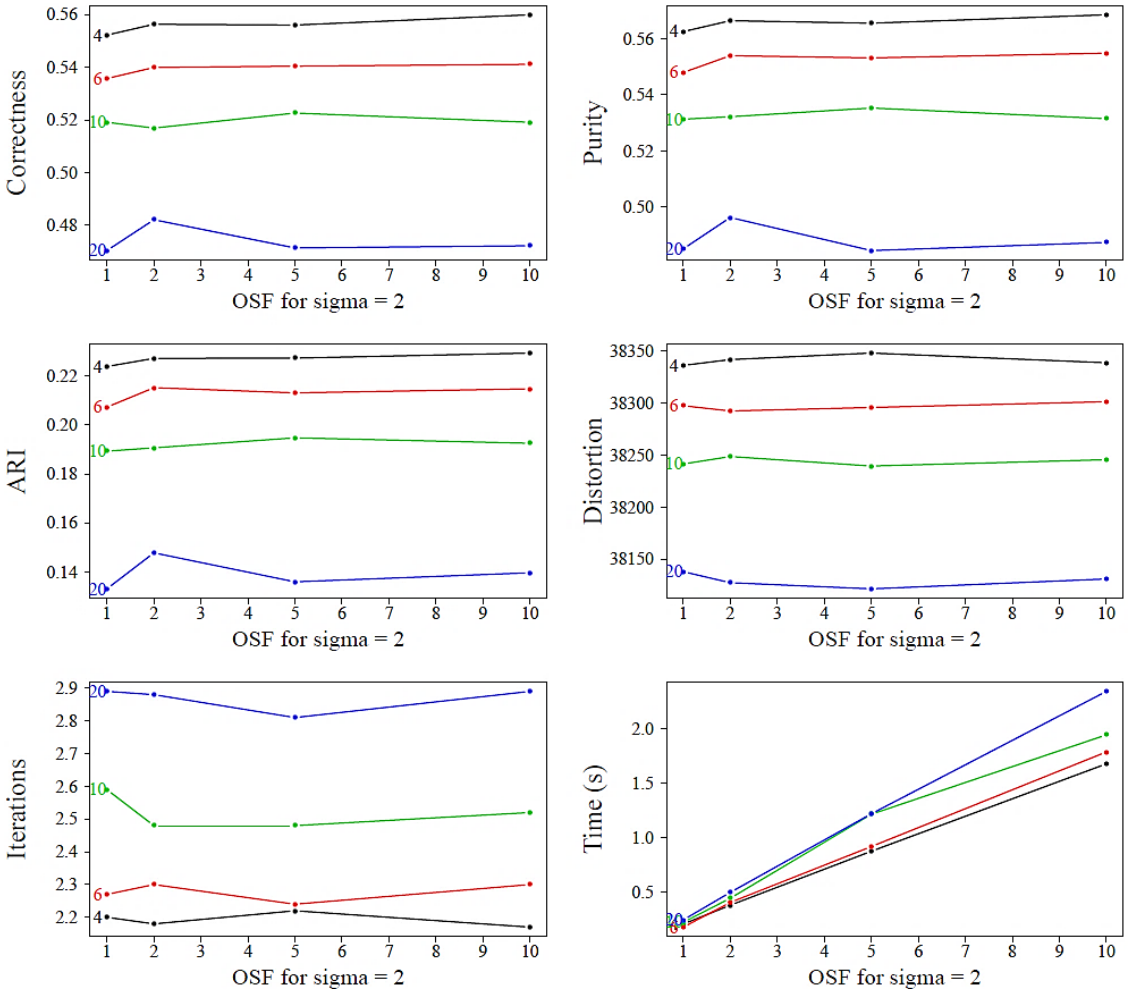


Fig. A.44. Model 1. OSF and DF behavior according to `sigma` = 2.

TABLE A.61.
MEAN & VARIANCE FOR MODEL 1. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 10.

`sigma` = 10

			OSF			
			1	2	5	10
Correctness Mean	DF	4	0.3421	0.3431	0.3433	0.3436
		6	0.3403	0.3399	0.3415	0.342
		10	0.3358	0.3339	0.3355	0.3368
		20	0.3345	0.3364	0.3355	0.3363
Correctness Variance						
	DF	4	0.0006067	0.0006863	0.0006244	0.0006677
		6	0.0007423	0.0008394	0.0008472	0.001034
		10	0.0006832	0.000725	0.0005785	0.0005291
		20	0.0006836	0.0005344	0.000724	0.0005306

			OSF			
			1	2	5	10
Purity Mean	DF	4	0.3497	0.3511	0.3508	0.3514
		6	0.3468	0.348	0.349	0.3509
		10	0.3435	0.342	0.3435	0.3441
		20	0.3429	0.3434	0.3439	0.3436
Purity Variance						
	DF	4	0.0006534	0.0007008	0.0005812	0.0006889
		6	0.0006846	0.0007192	0.0007626	0.0009376
		10	0.0007381	0.0007455	0.0006876	0.000582
		20	0.000641	0.0005499	0.0007089	0.0005687

			OSF			
			1	2	5	10
ARI Mean	DF	4	0.006645	0.007564	0.007606	0.007573
		6	0.00465	0.004909	0.00614	0.007662
		10	0.002065	0.001812	0.00172	0.001919
		20	0.001865	0.002675	0.0023	0.001598
ARI Variance						
	DF	4	0.0002331	0.0002732	0.0002449	0.0002868
		6	0.0002236	0.0002256	0.0002273	0.0003429
		10	0.000237	0.0001838	0.0001864	0.0001749
		20	0.0002429	0.0001877	0.0002939	0.0002121

			OSF			
			1	2	5	10
Distortion Mean	DF	4	952300	952400	952300	952400
		6	951100	950900	951100	950900
		10	948300	948200	948200	948400
		20	943700	943700	943900	944300
Distortion Variance						
	DF	4	217100000	226900000	225300000	219900000
		6	219900000	218200000	2.26e+08	209700000
		10	228500000	222500000	219200000	223400000
		20	222900000	220200000	205500000	218400000

			OSF			
			1	2	5	10
Iterations Mean	DF	4	2.25	2.23	2.2	2.25
		6	2.47	2.43	2.32	2.37
		10	2.57	2.58	2.56	2.55
		20	2.98	3.04	2.94	3.04
Iterations Variance						
	DF	4	0.1894	0.2193	0.2222	0.25
		6	0.3526	0.3486	0.2804	0.2557
		10	0.389	0.5693	0.3297	0.351
		20	0.6057	0.5438	0.6226	0.4832

			OSF			
			1	2	5	10
Time Mean	DF	4	0.1747	0.4283	0.859	1.461
		6	0.1913	0.3991	0.9465	1.796
		10	0.231	0.4	0.9481	1.864
		20	0.2109	0.4516	1.169	2.129
Time Variance						
	DF	4	0.02079	0.1422	0.4153	1.261
		6	0.0203	0.075	0.5434	1.706
		10	0.03073	0.07943	0.5203	1.682
		20	0.02337	0.08755	0.4817	1.829

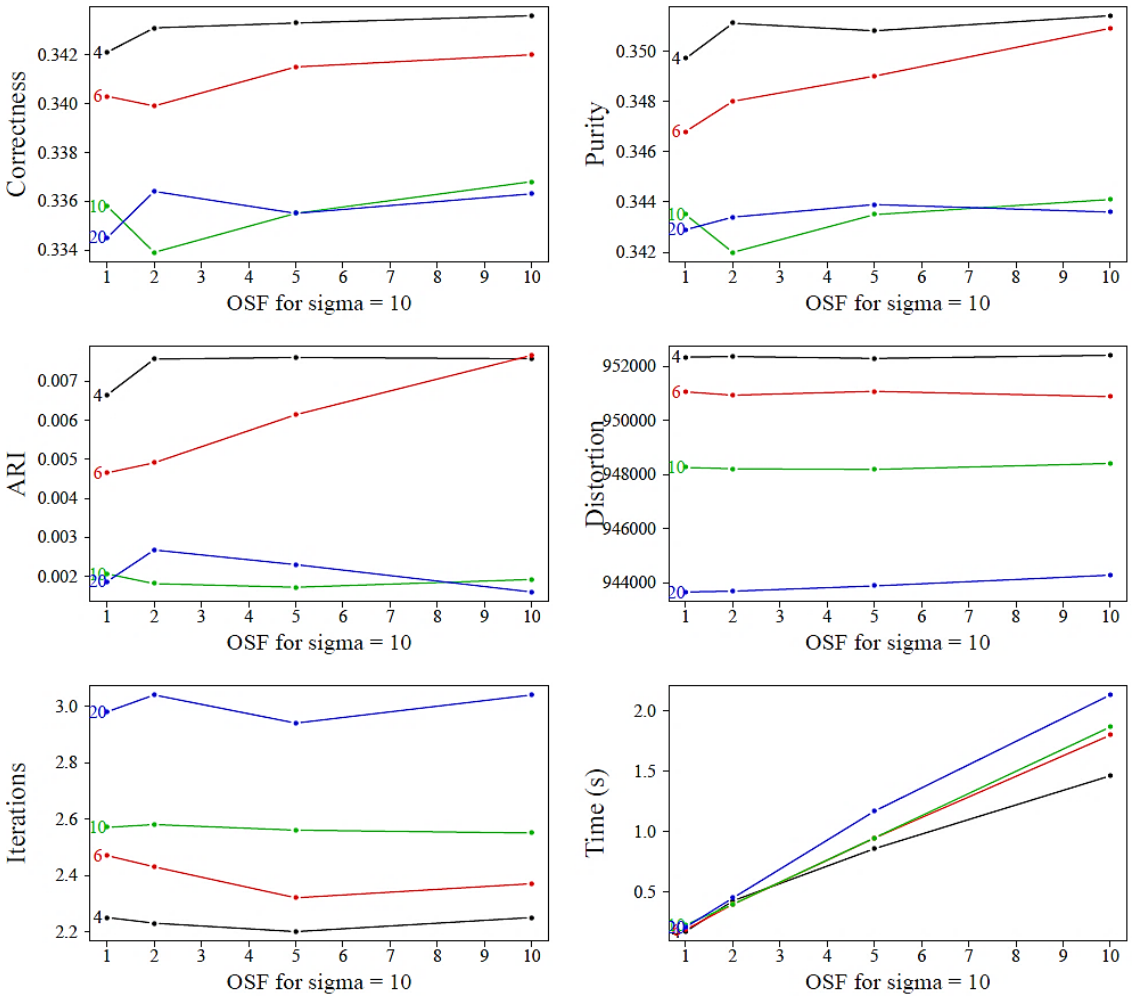


Fig. A.45. Model 1. OSF and DF behavior according to `sigma` = 10.

Model Two - OSF and DF Behavior According to Input Data (100 iterations)

TABLE A.62.
MEAN AND VARIANCE FOR MODEL 2. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 1.

`sigma` = 1

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Correctness	DF	4	0.989	0.9887	0.000102	0.0001064
		10	0.9991	0.9991	8,27E-03	8,27E-03
		15	0.9995	0.9996	4,80E-03	3,88E-03
		20	0.9994	0.9994	5,70E-03	5,70E-03

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Purity	DF	4	0.989	0.9887	0.000102	0.0001064
		10	0.9991	0.9991	8,27E-03	8,27E-03
		15	0.9995	0.9996	4,80E-03	3,88E-03
		20	0.9994	0.9994	5,70E-03	5,70E-03

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
ARI	DF	4	0.971	0.9702	0.0006766	0.0007061
		10	0.9976	0.9976	6,01E-02	6,01E-02
		15	0.9987	0.9989	3,48E-02	2,82E-02
		20	0.9984	0.9984	4,14E-02	4,14E-02

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Distortion	DF	4	9711	9711	21470	21280
		10	9687	9687	21600	21600
		15	9687	9687	21600	21630
		20	9687	9687	21610	21610

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Iterations	DF	4	1.38	1.38	0.238	0.238
		10	1.03	1.04	0.02939	0.03879
		15	1.02	1.01	0.0198	0.01
		20	1.03	1.01	0.02939	0.01

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Time	DF	4	0.06661	0.3101	0.0006209	0.003635
		10	0.06884	0.2978	0.0001583	0.0006878
		15	0.06896	0.3047	0.0003557	0.001364
		20	0.05759	0.2648	9,65E-02	5,55E-02

TABLE A.63.
MEAN & VARIANCE FOR MODEL 2. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 1.5.

sigma = 1.5

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Correctness	DF	4	0.9343	0.9337	0.0006066	0.0006377
		10	0.9748	0.9754	0.0002636	0.0002695
		15	0.9789	0.9786	0.0002099	0.0002162
		20	0.977	0.9771	0.0002414	0.0002309

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Purity	DF	4	0.9343	0.9337	0.0006066	0.0006377
		10	0.9748	0.9754	0.0002636	0.0002695
		15	0.9789	0.9786	0.0002099	0.0002162
		20	0.977	0.9771	0.0002414	0.0002309

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
ARI	DF	4	0.8383	0.8372	0.00302	0.003161
		10	0.934	0.9355	0.001671	0.001747
		15	0.9444	0.9437	0.001386	0.001404
		20	0.9397	0.9399	0.001543	0.001481

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Distortion	DF	4	21890	21890	114700	114300
		10	21790	21790	110200	109700
		15	21780	21780	109200	109600
		20	21780	21780	109700	109700

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Iterations	DF	4	1.64	1.6	0.2327	0.2626
		10	1.54	1.4	0.2509	0.2424
		15	1.55	1.51	0.25	0.2524
		20	1.6	1.56	0.2424	0.2489

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Time	DF	4	0.0665	0.3418	0.001375	0.05697
		10	0.07067	0.3007	0.0007388	0.003463
		15	0.07934	0.3405	0.003108	0.07882
		20	0.05854	0.2989	0.0001574	0.02755

TABLE A.64.
MEAN AND VARIANCE FOR MODEL 2. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 2.

`sigma` = 2

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Correctness	DF	4	0.8466	0.8474	0.00213	0.001779
		10	0.9108	0.9128	0.0009286	0.0008749
		15	0.9125	0.9139	0.0007664	0.0007715
		20	0.904	0.9066	0.001384	0.001061

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Purity	DF	4	0.8473	0.8474	0.001915	0.001779
		10	0.9108	0.9128	0.0009286	0.0008749
		15	0.9125	0.9139	0.0007664	0.0007715
		20	0.9041	0.9066	0.001346	0.001061

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
ARI	DF	4	0.6589	0.6589	0.005622	0.005296
		10	0.7794	0.7842	0.00472	0.004473
		15	0.7832	0.7863	0.00383	0.003893
		20	0.7665	0.7714	0.005059	0.003817

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Distortion	DF	4	38860	38850	357700	353200
		10	38670	38670	343900	346400
		15	38650	38650	349900	350000
		20	38640	38640	351100	351100

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Iterations	DF	4	1.93	1.94	0.1062	0.2388
		10	1.91	1.93	0.1029	0.1062
		15	2.01	1.93	0.07061	0.06576
		20	2.08	2.12	0.1147	0.1471

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Time	DF	4	0.1142	0.5034	0.0113	0.1838
		10	0.09298	0.3997	0.006125	0.1055
		15	0.1054	0.4913	0.012	0.1672
		20	0.1088	0.4551	0.0142	0.1669

TABLE A.65.
MEAN & VARIANCE FOR MODEL 2. OSF AND DF BEHAVIOR ACCORDING TO SIGMA = 10.

sigma = 10

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Correctness	DF	4	0.3739	0.381	0.0009311	0.000999
		10	0.3612	0.3582	0.001098	0.0008876
		15	0.354	0.3549	0.000996	0.000801
		20	0.3548	0.3527	0.0009626	0.001054

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Purity	DF	4	0.3845	0.3906	0.0009604	0.0009613
		10	0.3719	0.3692	0.001108	0.0009569
		15	0.3632	0.3633	0.0009977	0.0008668
		20	0.3642	0.3619	0.0009822	0.0009671

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
ARI	DF	4	0.03343	0.03859	0.0005101	0.0005317
		10	0.02408	0.02102	0.0005844	0.0004545
		15	0.01717	0.01705	0.00049	0.0003943
		20	0.01535	0.01692	0.0004857	0.000494

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Distortion	DF	4	953400	953400	218600000	216800000
		10	950300	950400	226200000	219400000
		15	947900	947600	209900000	211500000
		20	945400	946000	222400000	2.3e+08

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Iterations	DF	4	2.23	2.23	0.2395	0.2193
		10	2.65	2.51	0.3914	0.3938
		15	2.72	2.82	0.4057	0.4521
		20	3.06	2.9	0.7438	0.4343

			Mean		*Variance*	
			OSF		OSF	
			1	5	1	5
Time	DF	4	0.2191	0.9339	0.0304	0.5802
		10	0.2639	1.274	0.03966	0.6049
		15	0.2423	1.118	0.02868	0.6225
		20	0.2478	1.289	0.02859	0.4598

Model Three - DF Behavior According to Input Data (N = 100)

TABLE A.66.
MEAN AND VARIANCE FOR MODEL 3 PADI MEASURES. DF BEHAVIOR ACCORDING TO SIGMA = 0.5.

`sigma` = 0.5

Purity		*Mean*	*Variance*	*Variance*	*Mean*		*ARI*
		OSF = 1		OSF = 1			
Degrees of Freedom (DF)	10	0.769	0.003731	0.004439	0.5759	10	*Degrees of Freedom (DF)*
	11	0.8144	0.003865	0.00634	0.6387	11	
	12	0.7914	0.003999	0.005561	0.607	12	
	13	0.8175	0.004895	0.007475	0.6492	13	
	14	0.8102	0.003947	0.006405	0.6344	14	
	15	0.8179	0.003463	0.005912	0.6422	15	
	16	0.8098	0.004565	0.006754	0.636	16	
	17	0.7995	0.004848	0.007114	0.6228	17	
	18	0.7972	0.005321	0.007932	0.6233	18	
	19	0.7956	0.004664	0.006473	0.6176	19	
	20	0.7877	0.005383	0.006831	0.6124	20	

Distortion		*Mean*	*Variance*	*Variance*	*Mean*		*Iterations*
		OSF = 1		OSF = 1			
Degrees of Freedom (DF)	10	6026	3660	0.1269	2.12	10	*Degrees of Freedom (DF)*
	11	6020	3547	0.07434	2.08	11	
	12	6022	3617	0.1546	2.13	12	
	13	6019	3627	0.1147	2.08	13	
	14	6019	3639	0.1894	2.15	14	
	15	6018	3582	0.1111	2.1	15	
	16	6018	3653	0.08273	2.09	16	
	17	6019	3645	0.09091	2.1	17	
	18	6018	3639	0.2945	2.22	18	
	19	6017	3539	0.3292	2.21	19	
	20	6017	3678	0.3257	2.24	20	

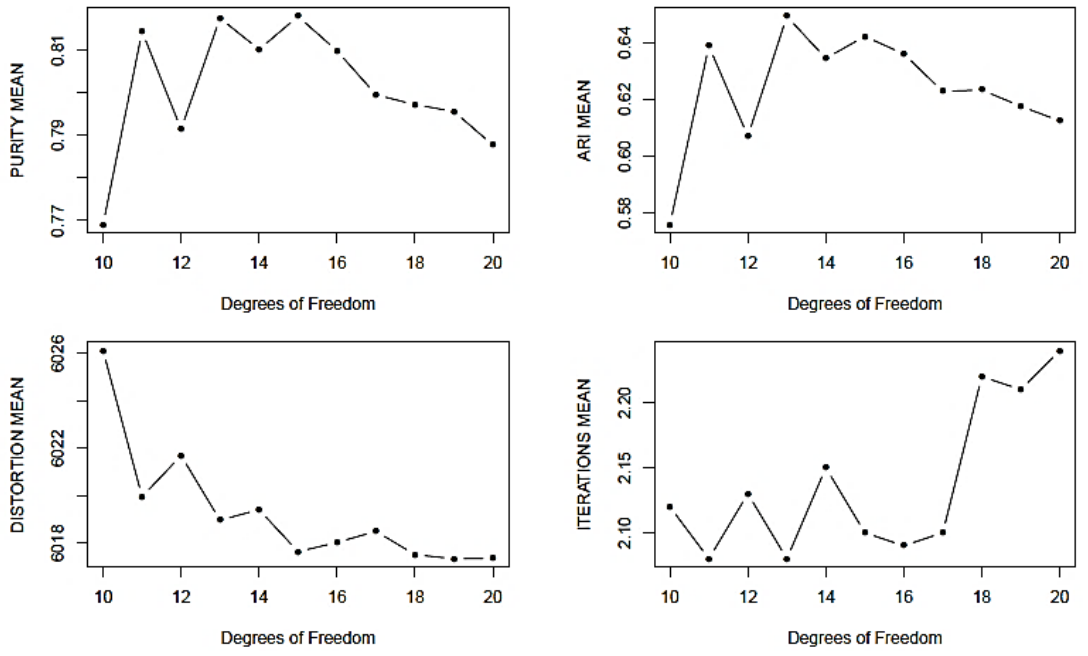


Fig. A.46. Model 3 PADI measures plot for `sigma` = 0.5.

TABLE A.67.
MEAN AND VARIANCE FOR MODEL 3 PADI MEASURES. DF BEHAVIOR ACCORDING TO SIGMA = 1.

`sigma` = 1

Purity	*Mean*	*Variance*	*Variance*	*Mean*	*ARI*
Degrees of Freedom (DF)	OSF = 1		OSF = 1		*Degrees of Freedom (DF)*
10	0.558	0.001247	0.00152	0.3267	10
11	0.5572	0.002171	0.002142	0.322	11
12	0.5523	0.001607	0.001545	0.3224	12
13	0.561	0.001708	0.00185	0.3317	13
14	0.5486	0.001741	0.001658	0.3214	14
15	0.5551	0.001255	0.001501	0.3206	15
16	0.5503	0.00149	0.002318	0.3236	16
17	0.5435	0.001613	0.001944	0.3192	17
18	0.5457	0.001395	0.001936	0.3174	18
19	0.5426	0.001782	0.001964	0.3174	19
20	0.5385	0.001679	0.002241	0.3143	20

Distortion	*Mean*	*Variance*	*Variance*	*Mean*	*Iterations*
Degrees of Freedom (DF)	OSF = 1		OSF = 1		*Degrees of Freedom (DF)*
10	2399	5570	0.3656	2.41	10
11	2399	5695	0.4541	2.52	11
12	2399	5811	0.351	2.55	12
13	2398	5847	0.507	2.59	13
14	2398	5813	0.4279	2.58	14
15	2398	5923	0.3688	2.57	15
16	2397	5971	0.486	2.67	16
17	2397	5539	0.4747	2.7	17
18	2397	5404	0.3979	2.69	18
19	2397	5568	0.4562	2.78	19
20	2396	5684	0.4671	2.76	20

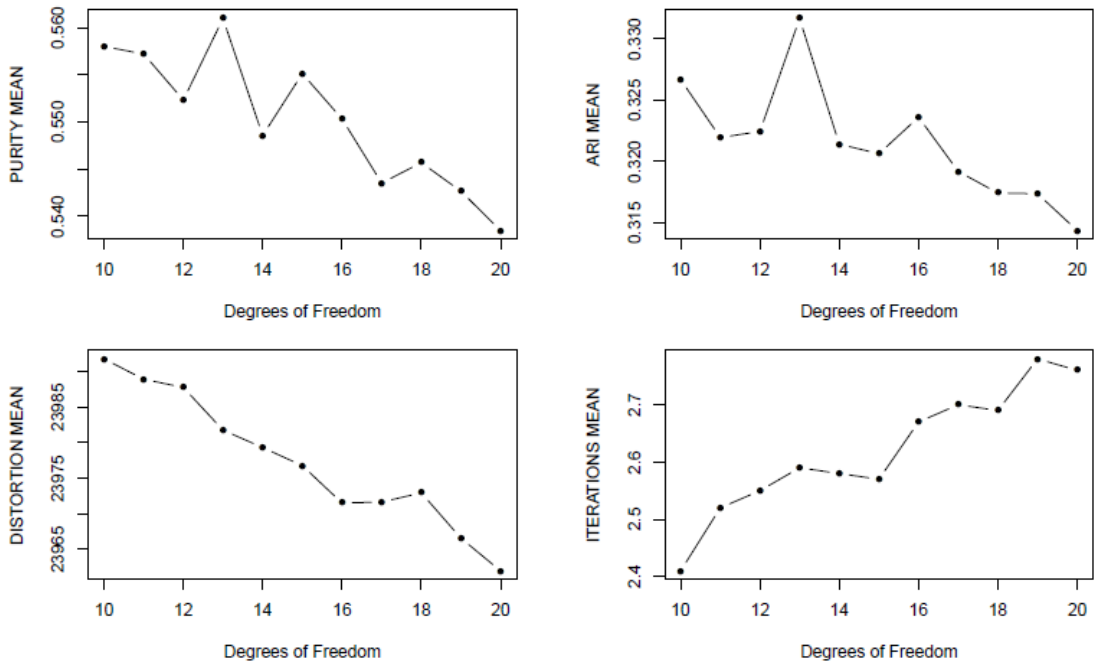


Fig. A.47. Model 3 PADI measures plot for `sigma` = 1.

TABLE A.68.
MEAN AND VARIANCE FOR MODEL 3 PADI MEASURES. DF BEHAVIOR ACCORDING TO SIGMA = 1.5.

sigma = 1.5							
		Mean	*Variance*	*Variance*	*Mean*		
Purity		OSF = 1		OSF = 1			*ARI*
	Degrees of Freedom (DF)					*Degrees of Freedom (DF)*	
	10	0.4616	0.001034	0.001191	0.1964	10	
	11	0.4621	0.001634	0.001495	0.1988	11	
	12	0.4562	0.001228	0.001292	0.1924	12	
	13	0.4592	0.001126	0.001465	0.196	13	
	14	0.457	0.001305	0.001716	0.1938	14	
	15	0.4567	0.001232	0.001487	0.1955	15	
	16	0.4556	0.001093	0.001512	0.1934	16	
	17	0.4548	0.001115	0.001095	0.1904	17	
	18	0.4506	0.001191	0.00118	0.1847	18	
	19	0.4494	0.001076	0.001465	0.1856	19	
	20	0.4476	0.0008952	0.001089	0.1849	20	

		Mean	*Variance*	*Variance*	*Mean*		
Distortion		OSF = 1		OSF = 1			*Iterations*
	Degrees of Freedom (DF)					*Degrees of Freedom (DF)*	
	10	5385	2928	0.3539	2.64	10	
	11	5385	2788	0.4873	2.76	11	
	12	5384	2832	0.406	2.91	12	
	13	5383	2995	0.3257	2.76	13	
	14	5381	2824	0.3939	2.7	14	
	15	5380	2905	0.4625	2.89	15	
	16	5380	2975	0.4948	2.99	16	
	17	5380	2919	0.7001	2.87	17	
	18	5379	3046	0.5708	3.07	18	
	19	5376	2928	0.4444	3	19	
	20	5376	2942	0.7373	2.99	20	

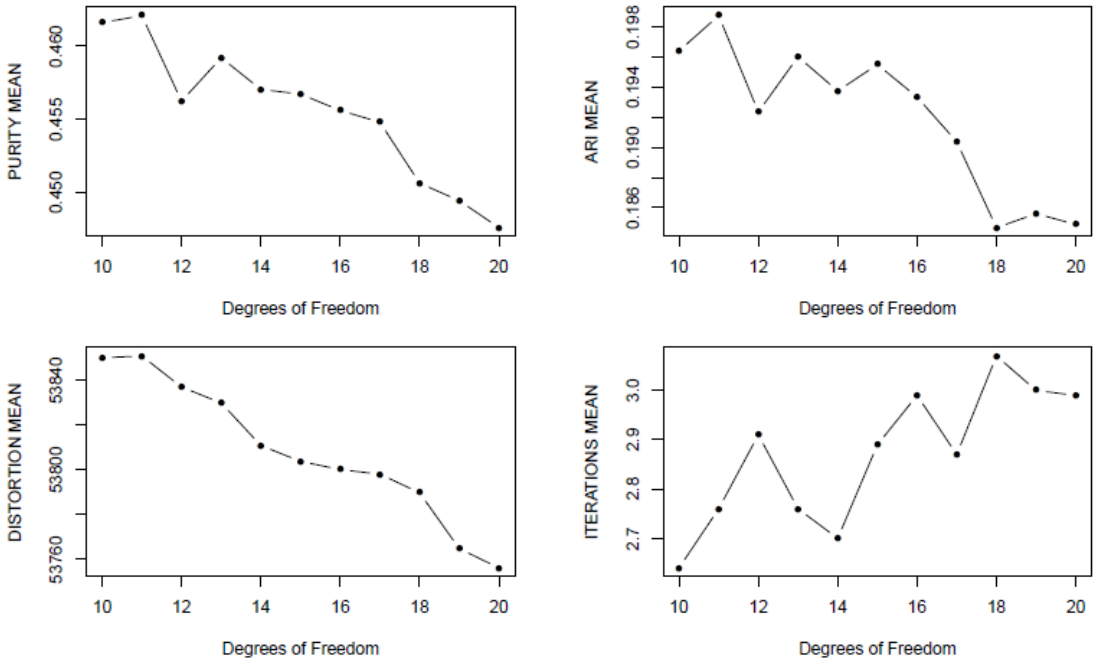


Fig. A.48. Model 3 PADI measures plot for sigma = 1.5.

TABLE A.69.
MEAN AND VARIANCE FOR MODEL 3 PADI MEASURES. DF BEHAVIOR ACCORDING TO SIGMA = 2.

`sigma` = 2

Purity		*Mean*	*Variance*	*Variance*	*Mean*		*ARI*
		OSF = 1		OSF = 1			
Degrees of Freedom (DF)	10	0.4126	0.0008898	0.1259	0.00111	10	*Degrees of Freedom (DF)*
	11	0.4098	0.001153	0.1221	0.001142	11	
	12	0.3996	0.001172	0.1169	0.001174	12	
	13	0.4031	0.001187	0.1178	0.001003	13	
	14	0.4044	0.001359	0.1155	0.001053	14	
	15	0.4019	0.0009737	0.1172	0.001071	15	
	16	0.4001	0.0008035	0.1115	0.0007753	16	
	17	0.399	0.000796	0.1114	0.0007851	17	
	18	0.3995	0.0008415	0.1092	0.0008382	18	
	19	0.3974	0.001074	0.1093	0.0009832	19	
	20	0.3975	0.001021	0.1083	0.001057	20	

Distortion		*Mean*	*Variance*	*Variance*	*Mean*		*Iterations*
		OSF = 1		OSF = 1			
Degrees of Freedom (DF)	10	9561	9139	0.2954	2.74	10	*Degrees of Freedom (DF)*
	11	9555	8697	0.5466	2.83	11	
	12	9556	8724	0.4375	2.87	12	
	13	9556	9437	0.4334	2.97	13	
	14	9553	9351	0.5732	2.95	14	
	15	9549	9218	0.7842	3.06	15	
	16	9547	8868	0.6206	3.16	16	
	17	9546	8797	0.4646	3	17	
	18	9545	8917	0.4469	3.24	18	
	19	9543	9183	0.644	3.32	19	
	20	9540	8786	0.654	3.15	20	

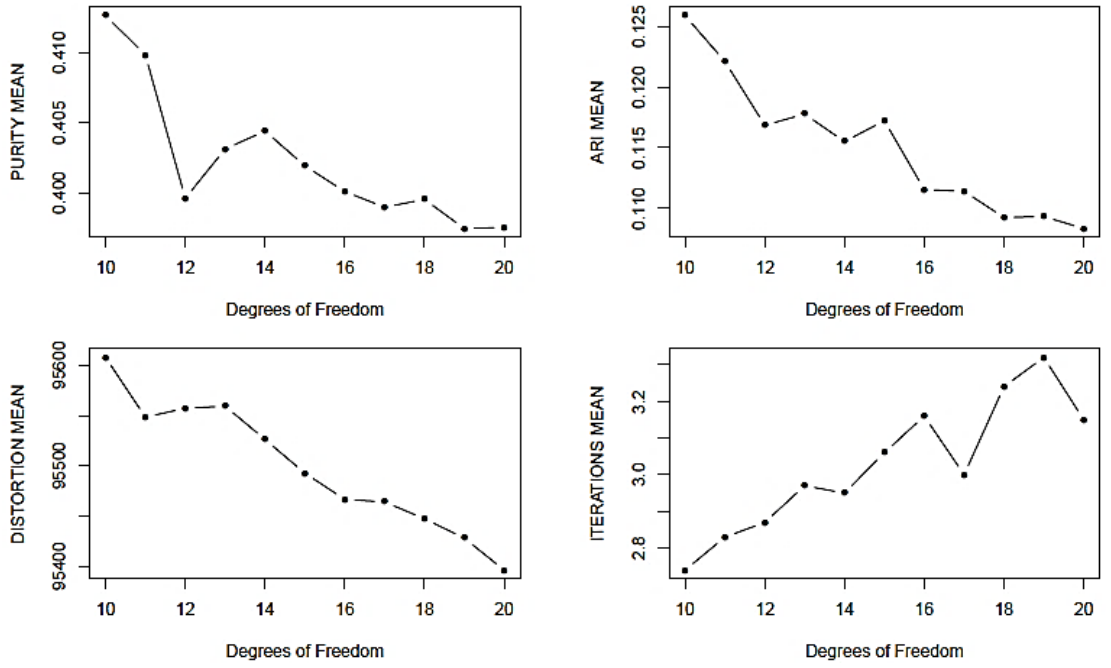


Fig. A.49. Model 3 PADI measures plot for `sigma` = 2.